Chain of Debate

Omer Abbas, Liam May, Kovidh Jain

01

Introduction

Background

- AI is everywhere
  - Rise of LLMs
- Benefits
  - Improved productivity
  - Efficient access of knowledge
  - Non-stop tutors
- Drawbacks
  - Misinformation
    - Deceptive confidence
    - Hallucinations
      - Not just deception but contrived facts and events

Limitations in Current Approaches

Existing approaches

- Exclusion of Misinformation from LLM Training Data
    - Hard to do without a model that can detect misinformation
- LLM Reinforcement Learning w/ Human Feedback (RLHF)
    - Misinformation can cause a topic to be controversial, even though there's a true answer.
        - Hard for humans to judge
    - Hard to adapt to new information
- Content filters
    - Impossible to create an exhaustive list

Intuition

- Configuring a model to output the correct response:
  - Is there a correct response?
  - Who decides that?
  - Does it change with evolving information?
- Alternative → Assume the correct response is unknown on query:
  - Explore the possibility space of every stance that could be taken
  - Through debate, expose the faulty logic
    - The best answer is the most substantiated and survives the pressure of being challenged
  - Mirrors ensemble learning except for ideas
    - The sum of all independent ideas is less likely to be wrong

Hypothesis

Having models deliberate on a prompt through Chain of Debate before issuing a final response will improve the quality of responses

02

Related Works

Papers

- Chain of Agents: Large Language Models Collaborating on Long-Context Tasks
    - Addresses the limitations of the context window in LLM's
        - Vanishing problem after prolonged sessions of prompting
        - Proposal: sub-contexting using agents
- MAD-Sherlock: Multi-Agent Debates for Out-of-Context Misinformation Detection
    - Cross validation from multiple agents to handle misleading prompts references
        - Image that's not consistent with the description
        - Also references sub-contexting using agents
- Exploring the Role of Large Language Models in Fake News Detection
    - Using LLM's to propose multi-perspective narratives for fake news detection
        - Incorporation of the proposal is handled to SLM's

Papers

- VeraCT Scan: Retrieval-Augmented Fake News Detection
    - Using a RAG to distill the facts and then scours the internet for conflicting or supporting evidence
- An Empirical Analysis on Large Language Models in Debate Evaluation
    - Limitations of using LLM's debate evaluators
        - Analyzes biases in judgments
            - GPT-3.5 and GPT-4 favor the second-to-speak and concluding speaker
- CrAM: Credibility-Aware Attention Modification in LLMs for Combating Misinformation In RAG
    - Proposes a method for weighing the credibility of sources for RAGS
        - Less credible sources have less influence

03

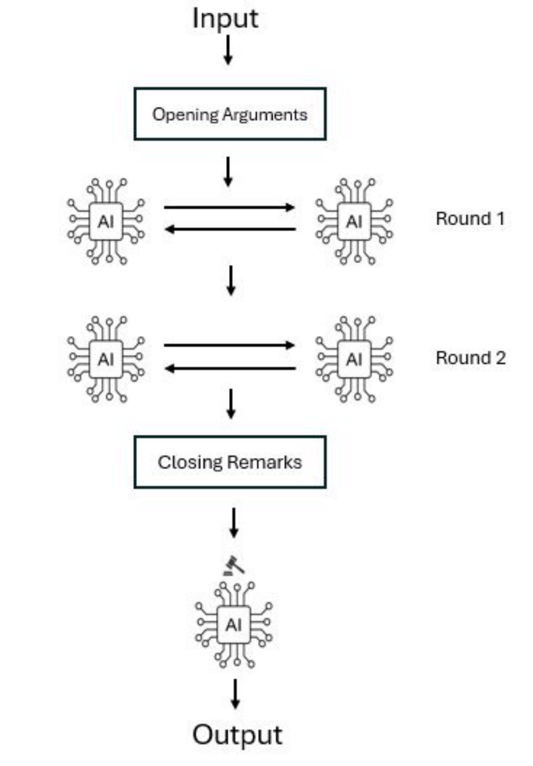
Methodology

Setup

- Initialization
  - Initialize debaters as independent models
    - Take opposing stances on an issue
      - True news of fake news
- Debate
  - Models interact with each following a debate format
- Verdict
  - Judge model selects best argument as a final response: Fake or True
- Dataset
  - Fake News Detection Dataset
    - Data: title, subject, and content of an article
    - Labels: Boolean
      - Whether the article is fake news
    - Take 1500 (out of 44898) subset without replacement at random for testing
  - FEVER
    - WIP - scraping restrictions

Debate Format

- Structure
  - Opening arguments
    - Initial arguments to set the stage for the debate
  - 2 rounds of responses
    - Round 1
      - Opening argument responses for each model
    - Round 2
      - Counter responses to round 1 arguments
  - Closing remarks
    - Highlights of strongest arguments for each model



Metrics

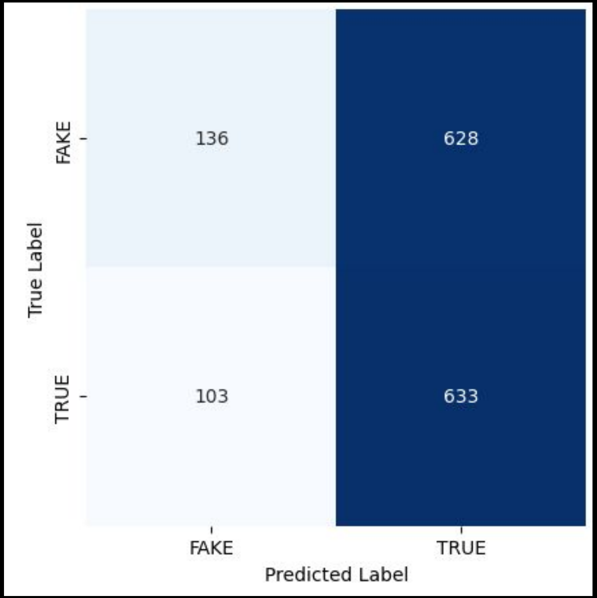
- Accuracy
  - Correct predictions out of all the training examples
- Precision
  - Reliability of a model in its relevant class predictions
    - Penalizes mistaking a fake news article for a true one
    - Measured as the ratio of true positives over true positives and false positives
- Recall
  - Reliability of a model among all relevant class examples
    - Penalizes mistaking a true news article for a fake one
    - Ratio of true positives and true positives + false negatives
- F1
  - Average of precision and recall
- Confusion matrix

04

Experiment Results and Analysis

Results - Base Model

- 1500 examples
    - 764 are fake
    - 736 are true
    - Positive = true articles
- 136 true negatives (top left)
    - Correctly identified fake articles
- 103 false negatives (bottom left)
    - Mistaking true articles for fake ones
- 628 false positives (top right)
    - Mistaking fake articles for true ones
- 633 true positives (bottom right)
    - Correctly identified true articles



Results - Base Model

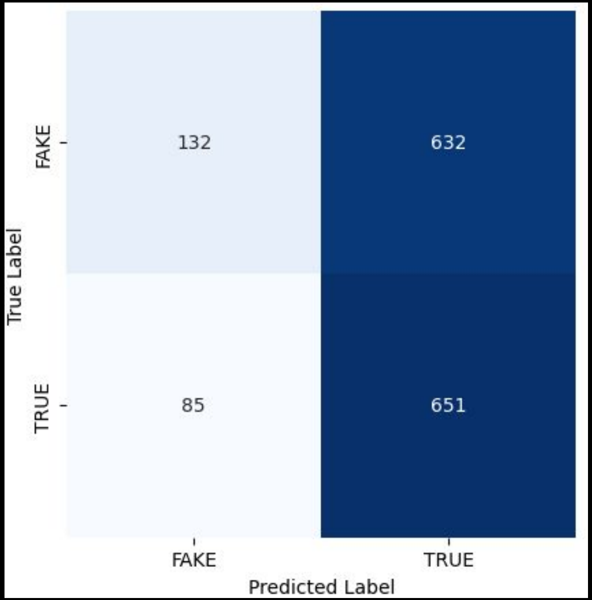
- 51.26% Accuracy
- 53.6% weighted average precision
    - Number of examples as a weight
- 51.3% weighted average recall
- 44.92% weighted average F1

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

Results - Chain of Debate

- Confusion matrix
  - 132 true negatives (-6)
    - Fewer correct predictions of fake articles
  - 85 false negatives (-15)
    - Reduction in mistakes for true examples
  - 632 false positives (+4)
    - Slight increase in mistakes of fake examples
  - 651 true positives (+18)
    - More correct classifications of true examples
- Net result:
  - (15 + 18) - (6 + 4) = 21
    - A better confusion matrix



Results - Chain of Debate

- 52.2% accuracy (+.93%)
- 55.9% avg precision (+2.3%)
- 52.2% avg recall (+.9%)
- 45.4% avg F1 (+.5%)
- Net improvement!

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

Patterns

- General
  - Recall
    - Both experiments yielded poor recall for fake instances
    - Performant on recall for true
  - Precisions
    - Comparable but stronger performance for fake instances
- Insight
  - If an article is indeed true, it will likely be detected by the model significantly more than if it were fake
  - A model is less likely to misclassify fake news article than a true news article

Base Model Metrics

	Precision	Recall
FAKE	0.569	0.178
TRUE	0.502	0.86

Chain of Debate

	Precision	Recall
FAKE	0.608	0.173
TRUE	0.507	0.884

Comparison to Simple Models

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

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

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

05

Conclusion and Future Work

Effectiveness of Our Approach

- Post training optimizations
  - Fine-tuning
  - Debate framework restructuring
  - Allows for combining different models
- Can be used to detect misinformation AND generate responses
  - In contrast to non-generative models
- Easy to catch where the model went off track
  - Single-response models have no incentive to expand on their assumptions
    - Source of hallucinations

Future Work

- Fine-tuning
  - Misinformation
- Implementing Search
  - Saw initial improvements
  - Source Quality Analysis
- Integration of Encoder Only Models
  - Simplifies framework
  - Reduces Computation
- Improve Debate Frameworks
- Apply Improvements to Similar Benchmarks

Questions?

Works Cited

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. Cram: Credibility-aware attention modification in llms for combating misinformation in rag, 2024.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Badactor, good advisor: Exploring the role of large language models in fake news detection. Proceedings of the AAAI Conference on Artificial Intelligence, 38(20):22105–22113, March 2024.

Kumud Lakara, Georgia Channing, Juil Sock, Christian Rupprecht, Philip Torr, John Collomosse, and Christian Schroeder de Witt. Llm-consensus: Multi-agent debate for visual misinformation detection, 2025.

Xinyi Liu, Pinxin Liu, and Hangfeng He. An empirical analysis on large language models in debate evaluation, 2024.

Cheng Niu, Yang Guan, Yuanhao Wu, Juno Zhu, Juntong Song, Randy Zhong, Kaihua Zhu, Siliang Xu, Shizhe Diao, and Tong Zhang. Veract scan: Retrieval-augmented fake news detection with justifiable reasoning, 2024.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan ¨O. Arik. Chain of agents: Large language models collaborating on long-context tasks, 2024.

Image Sources

iStock. (n.d.). Artificial intelligence illustrations. iStock. https://www.istockphoto.com/illustrations/artificial-intelligence

IconScout. (n.d.). Database [Icon]. IconScout. https://iconscout.com/free-icon/database-827_444649

iStock. (n.d.). Gavel silhouette illustrations. iStock. https://www.istockphoto.com/illustrations/gavel-silhouette