Mitigating/Grading Mis-&Disinformation with Hybrid Model AI

David Sun [1] Eric Gu [1] Eric Sun [1] Mentor: Ali Arsanjani [2]

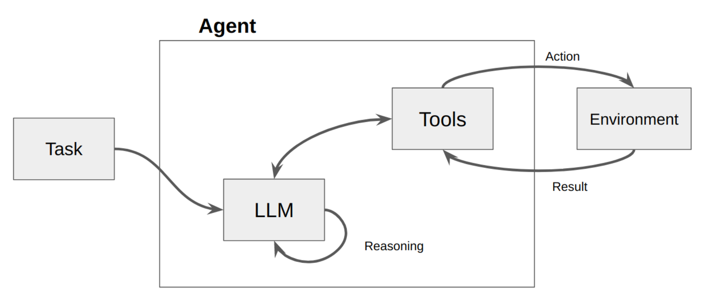
dasun@ucsd.edu ergu@ucsd.edu z9sun@ucsd.edu arsanjani@google.com

[1]University of California, San Diego [2]Google AI

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Background

This project introduces a hybrid veracity detection and scoring framework that leverages both generative AI and traditional machine learning to detect, rank, and mitigate misinformation and disinformation across diverse media formats. This hybridized LLM-based veracity machine not only facilitates precise misinformation detection but also provides a scalable and interpretable solution for managing the complexities of content veracity in an evolving digital landscape.



Datasets

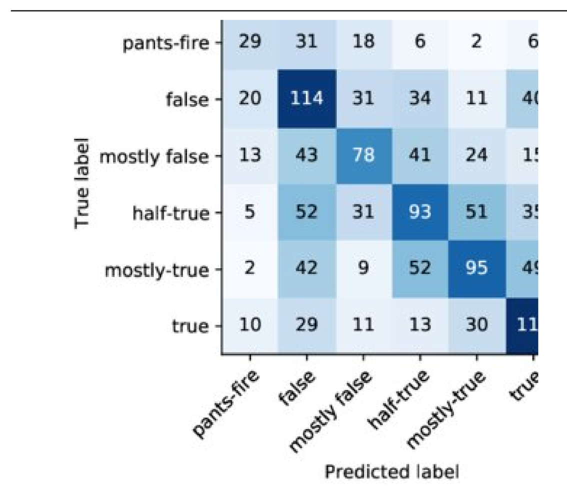


Figure 1. Confusion Matrix for Liar PLUS

Liar PLUS: Integration with Predictive AI

Feature Extraction: The predictive AI can analyze the text data from LIAR-PLUS to extract features that are relevant to determining the truthfulness of statements.

Contextual Analysis: The detailed justifications provided in LIAR-PLUS allow the predictive AI to learn not just whether a statement is true or false, but why it was categorized as such.

Training Data: LIAR-PLUS serves as training data for the predictive AI (random forest model).

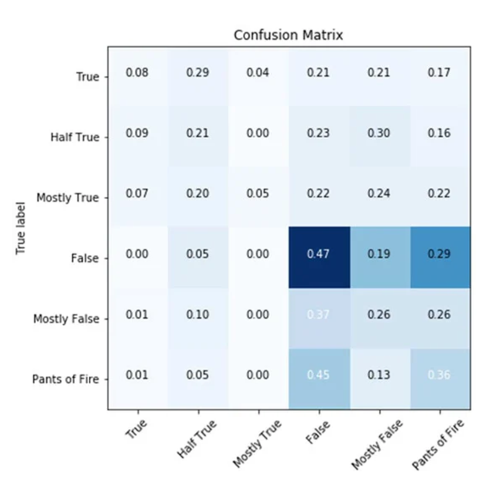


Figure 2. Confusion Matrix for Politifact

Politifact & Snopes: Integration with Generative AI

Integration of Ground Truth Labels: These datasets offer crucial ground truth labels, including "True," "Half-True," and "False," along with detailed explanations that elucidate the rationale behind these assessments.

Incorporation into System Prompts: This enhancement will enable the veracity machine to provide users not only with accurate classifications but also the reasoning behind these classifications.

Type	True	False
6-class	True: 12% Mostly True: 19% Half True: 21%	Mostly False: 18% False: 18% Pants-on-fire: 12%
4-class	Factual	Incomplete / Manipulative / Hoax
Statements	6,096	
# Shares	124,215	
# Comments	38,963	

Table 1. Distribution of Statements by Truthfulness Category in Politifact

Roadmap

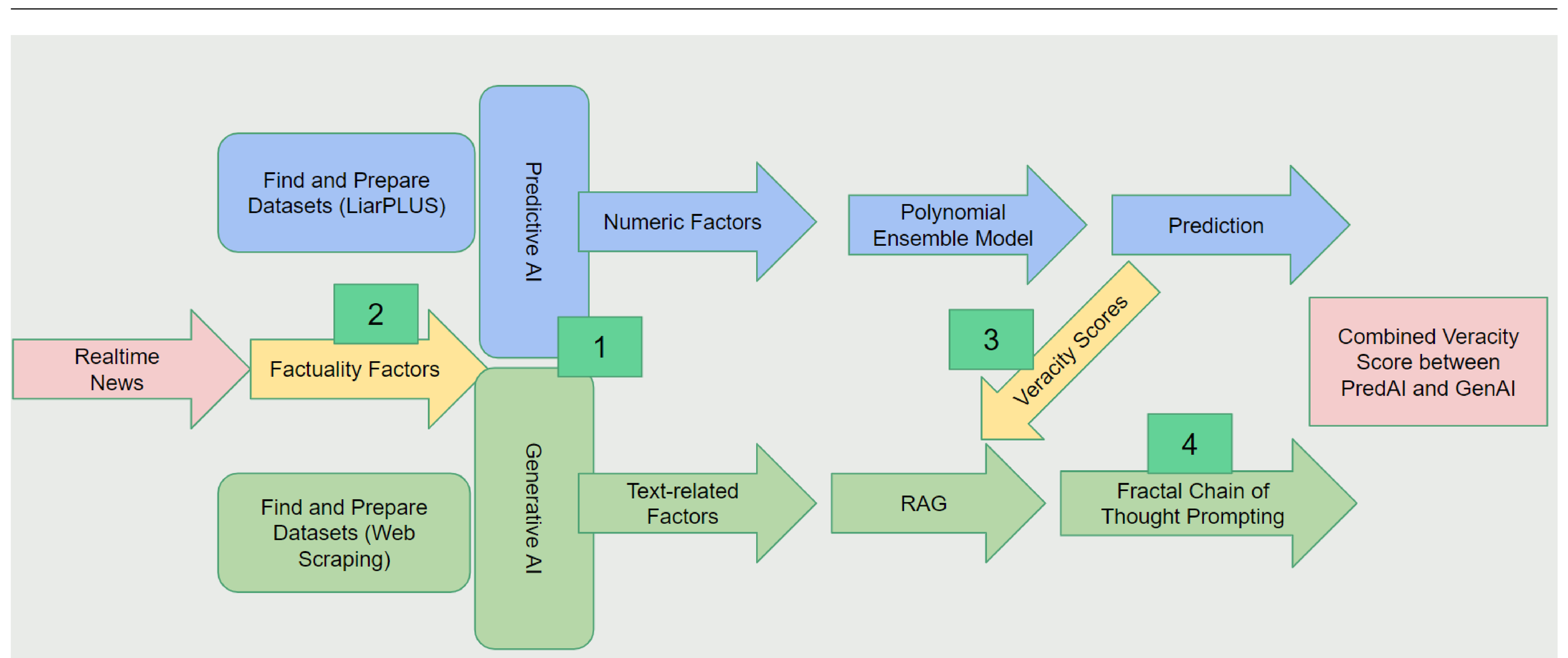


Figure 3. The flowchart of our project.

Methodology

Predictive AI

Combines traditional predictive AI models for statistical rigor and generative AI for nuanced content analysis. Anchors analysis with structured factuality scoring.

- **Dataset:** LiarPlus, fact-checking and fake news detection dataset
- **Factuality features:** Location, Education, Event coverage, Echo chamber, News coverage, Malicious account
- **Trained Model:** Random forest classifier
- **Output labels:** True, mostly-true, half-true, barely-true, false, pants-fire

Generative AI

- **Factuality Factors:** Content veracity assessed through multi-dimensional factors, enableing precise and transparent decomposition of misinformation.
- **Retrieval-Augmented Generation (RAG):** All relevant data is stored in ChromaDB, which acts as a retrieval-augmented generation (RAG) system for our model. ChromaDB enables us to organize and manage a vast collection of content fragments, which can be referenced by the AI to provide contextually accurate responses. This RAG system significantly improves the model's capability to handle complex misinformation scenarios by accessing precise data points in real-time.

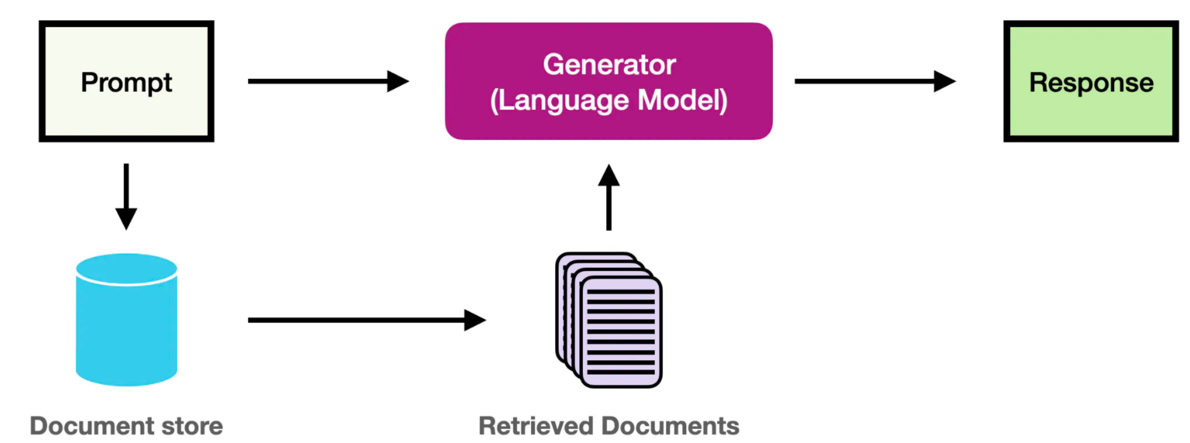


Figure 4. RAG Illustration

- **SerpAPI Web Search:** By embedding these real-time search results into the prompt, the GenAI gains access to a broader and more dynamic set of data, enabling it to cross-reference claims made in the inputted news article with credible external sources.

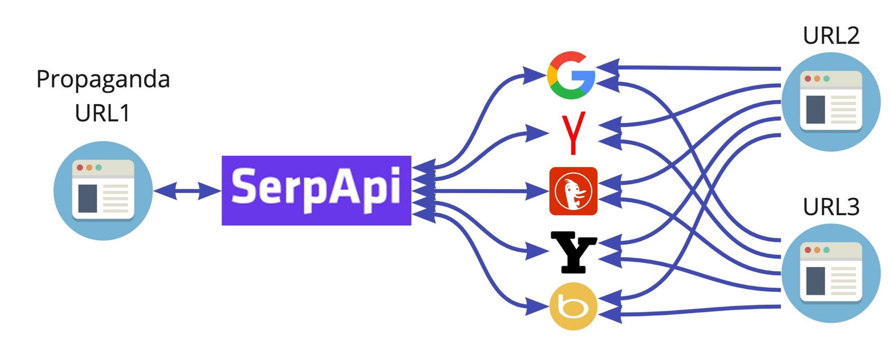


Figure 5. SerpAPI Illustration

- **Fractal Chain of Thought (FCOT) Prompting:** Advances traditional chain-of-thought prompting with iterative, layered analysis:
  - Evaluates factuality factors in multiple iterations.
  - Incorporates feedback loops for refined insights and improved veracity scoring.
- **Function Calling:** Function calls are strategically used to dynamically adjust analysis parameters based on real-time feedback. This adaptability is essential for calculating the effectiveness of various thought patterns generated by our algorithm, ensuring that the most logical and factually consistent chains are prioritized.

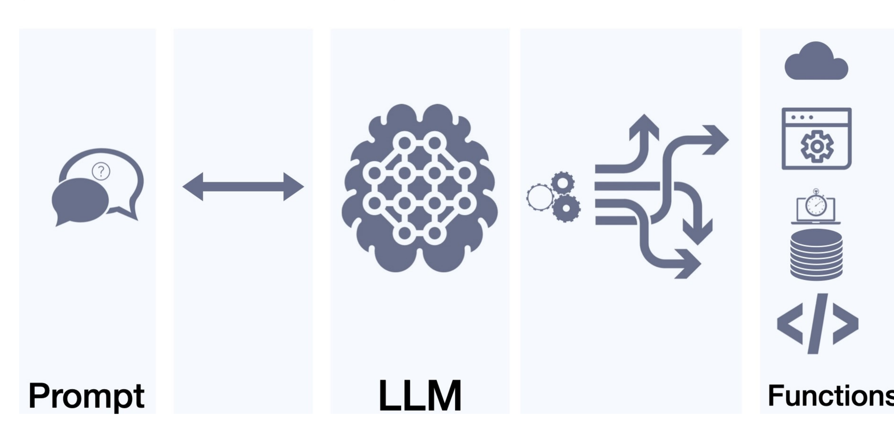


Figure 6. Function Calling Illustration

Together, these tools form a robust architecture where the generative AI system can ground its responses in fact-checked and contextually relevant data, providing a structured and rigorous approach to misinformation detection.

Result

Model Description	Score (%)
BERT Embedding Model	43.7
XGBoost/LightGBM (Boosting algorithm)	33.1
Random Forest Classifier (Bagging algorithm)	67.8
Sentiment Analysis (TF-IDF)	45.9
Word2Vec	55.2

Table 2. Predictive Performance on Liar PLUS dataset

Model Description	Score (%)
Baseline (Feeding straight into Gemini Flash 2.0)	19
Hybrid (Random Forest + Gemini)	34.3
Hybrid + RAG	40
Hybrid + RAG + Web Search	56.9
Hybrid + RAG + Web Search + FCOT Prompting	67.2
Hybrid (50/50) + RAG + Web Search + FCOT Prompting + Function Calling	65.3
Hybrid (70/30) + RAG + Web Search + FCOT Prompting + Function Calling	85.1

Table 3. Overall Model Performance

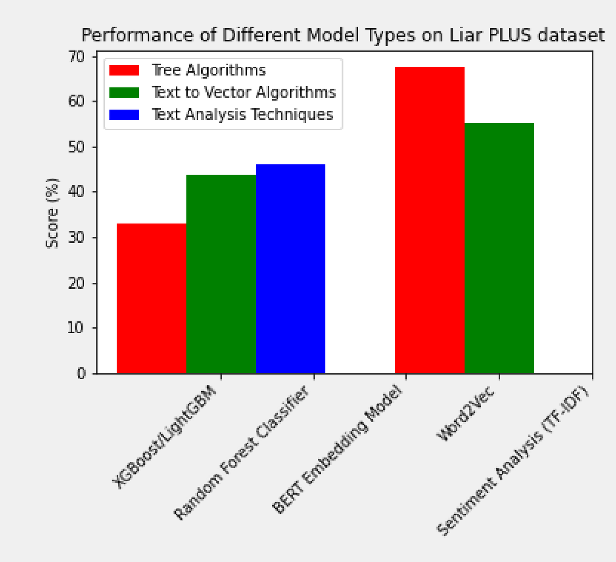


Figure 7. Predictive Algorithm Comparison

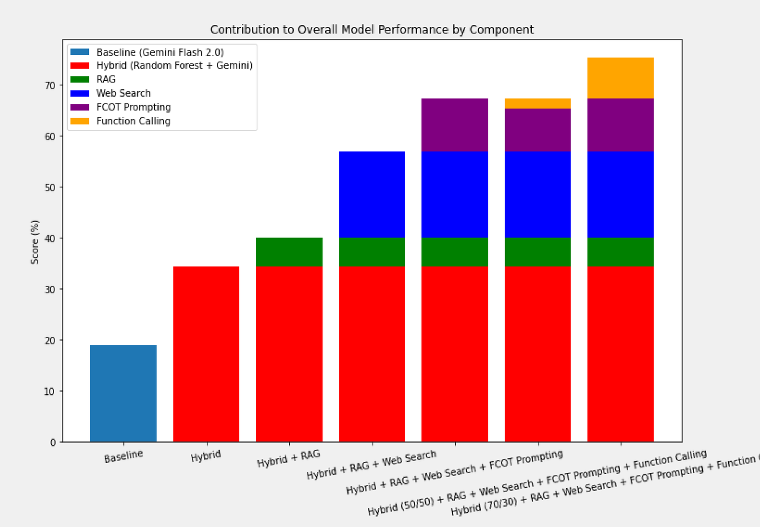


Figure 8. Stacked Bar Chart for overall performance

Discussion

Figure 9. Table for Prompting comparison

	Predictive		Generative		Hybrid	
Category	Precision	Recall	Precision	Recall	Precision	Recall
barely-true	0.17	0.12	0.12	0.05	0.13	0.10
FALSE	0.28	0.36	**0**	**0**	0.21	0.23
half-true	0.30	0.19	0.33	0.32	0.23	0.18
mostly-true	0.35	0.63	0.05	0.08	0.42	0.62
pants-fire	0.64	0.55	0.16	0.50	0.54	0.71
TRUE	0.32	0.14	0.16	0.16	0.37	0.17
Overall	0.30	0.31	0.15	0.17	0.30	0.31

Table 4. Performance comparison of Predictive, Generative, and hybrid models

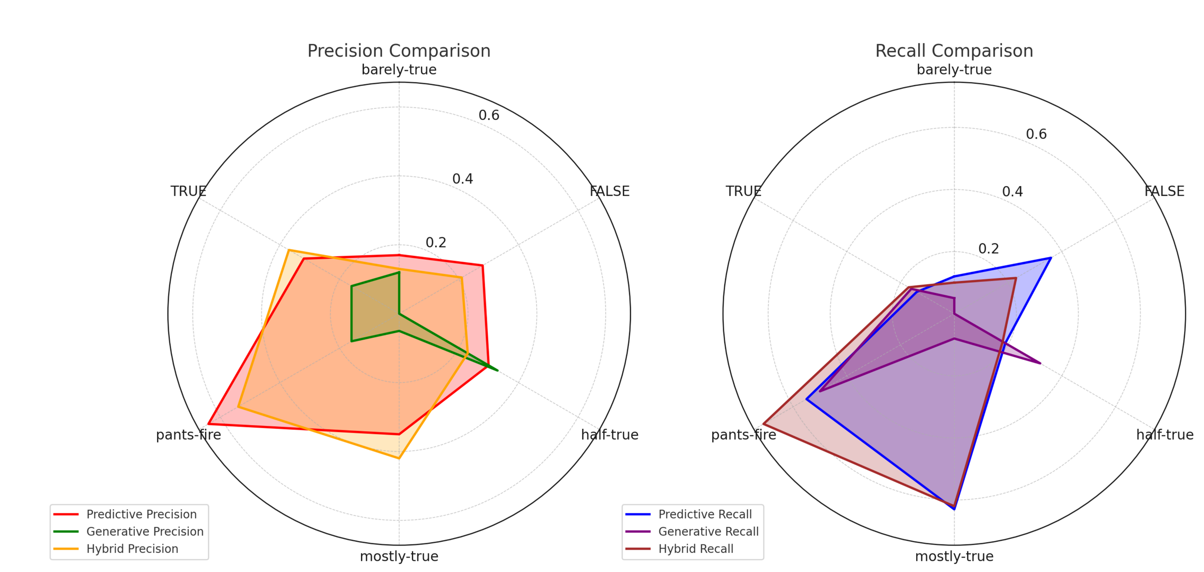


Figure 10. Radar Chart Comparison for Predictive, Generative, Hybrid Models.

Conclusion

Moving forward, we aim to expand our dataset and refine our algorithms to better handle the dynamic and evolving nature of online information. Future work will focus on automating the integration of real-time data feeds and enhancing the system's adaptability to new and emerging types of misinformation. We also plan to explore the ethical implications of AI in information verification, ensuring that our advancements in AI veracity technologies are aligned with societal values and norms.