Mitigating/Grading Misinformation and Disinformation with Hybrid Model AI

Eric Sun[†]
Halıcıoğlu Data Science Institute
University of California, San Diego
San Diego, California
z9sun@ucsd.edu

Eric Gu[†]
Halıcıoğlu Data Science Institute
University of California, San Diego
San Diego, California
ergu@ucsd.edu

David Sun[†]
Halıcıoğlu Data Science Institute
University of California, San Diego
San Diego, California
dasun@ucsd.edu

Dr. Ali Arsanjani
Research Advisor
Google Cloud
San Diego, California
arsanjani@google.com

Contents

1 Introduction 1
1.1 Related Works 2
1.2 Our Approach 2

2 Factor Usage 2
2.1 Factuality Factors 3

3 Methodologies 3
3.1 Predictive AI for Enhanced Misinformation Detection 4
3.2 Data Collection and Preprocessing Pipeline 4
3.3 External Datasets 4
3.4 External Search Engine 5
3.5 Function Calling 5
3.6 Prompting Techniques 5

4 Current results & outcomes 6
4.1 Current Stage 6
4.2 Model Performance Comparison . . 7
4.3 FCoT vs. Traditional CoT 7
4.4 Potential Improvements 8

5 Conclusion 8

6 Appendix 9

7 Contribution 9

References 9

Abstract—The proliferation of fake news and misinformation, often amplified by large language models (LLMs), poses a significant threat to societal trust and stability. This paper 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. Our approach decomposes content into structured analytical components, using an ensemble of factuality factors such as frequency heuristics, malicious account indicators, and psychological manipulation cues to identify and assess deceptive patterns. By employing advanced techniques such as Retrieval-Augmented Generation (RAG), fractal chain-of-thought prompting, and function calling, our system dynamically refines predictions, enhancing transparency and reducing hallucinations. 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.

1. Introduction

In today's digital era, the rapid spread of misinformation and disinformation poses a significant societal challenge. Enabled by the rise of advanced technologies such as large language models and artificial intelligence tools, these phenomena undermine mutual trust and can have serious consequences on democratic processes and public safety. Individuals and entities can now easily create and disseminate unchecked information, reaching vast audiences at an alarming rate.

This ease of spreading falsehoods not only threatens social harmony but also necessitates an urgent call for

1. [†]These authors contributed equally to this work

effective detection, evaluation, and mitigation strategies. This paper aims to explore the growing impact of digital misinformation and disinformation, highlighting how emerging technologies facilitate their spread. It will also propose new solutions to enhance the resilience of information ecosystems against the onslaught of digital falsehoods.

1.1. Related Works

The field of misinformation and disinformation detection has evolved significantly with the advent of Large Language Models (LLMs) and multi-modal approaches. Early detection systems primarily relied on traditional linguistic features and neural models [1], but recent research has shifted toward more sophisticated hybrid approaches that combine multiple modalities (text, images, social context) and leverage the reasoning capabilities of LLMs [2]. Studies have shown that simple feature concatenation across modalities is insufficient; instead, more nuanced approaches like attention mechanisms and logic-based reasoning (as demonstrated by LogicDM) have proven more effective [3]. The introduction of chain-of-thought prompting and guided LLMs has particularly revolutionized detection capabilities, outperforming traditional models like RoBERTa, especially on complex datasets [4]. Frameworks like SNIFFER have achieved significant breakthroughs in detecting out-of-context misinformation [5], while DISCO has demonstrated the importance of explainability in detection systems [6]. However, persistent challenges remain, including political bias in detection models, poor performance on novel events due to overfitting, and vulnerability to sophisticated cross-modal manipulations. Research has also expanded into agent-based frameworks for studying misinformation spread, highlighting the potential of LLMs in simulating realistic information ecosystems for testing countermeasures [7].

1.2. Our Approach

Our approach introduces a novel veracity scoring system that combines the reasoning capabilities of large language models with the reliability of traditional machine learning pipelines. Rather than treating news articles as monolithic units, we decompose them into meaningful chunks, enabling granular analysis of content veracity. This chunking strategy allows us to identify specific problematic sections within otherwise truthful content and provides a more nuanced understanding of how misinformation manifests within articles.

The core of our system employs a hybrid architecture that leverages both generative AI and traditional machine learning models. Our traditional predictive model, trained on an existing news detection dataset, LiarPlus by Tariq60, generate individual score predictions for various factuality factors, each representing different ways content can be misleading. These prediction scores serve as quantitative anchors for the analysis, providing statistical rigor that helps constrain the generative model's output space and reduce the risk of hallucination.

The system enhances these predictions through carefully constructed Retrieval-Augmented Generation (RAG), where similar statements are retrieved from a vector database of verified content. This retrieval process is guided not only by semantic similarity but also by crucial metadata such as temporal proximity, source credibility, and topic relevance, ensuring that the retrieved context is both accurate and pertinent to the analysis. By combining structured prediction scores with contextually relevant retrieved statements, our system creates a robust foundation for veracity assessment that maintains factual grounding while leveraging the analytical capabilities of generative AI.

We introduce a novel prompting technique called fractal chain-of-thought, which guides our generative model through an iterative process of defining and refining an objective function based on the identified factuality factors. This technique represents an advancement over traditional chain-of-thought prompting, enabling more transparent and structured reasoning paths for verification. The iterative refinement process allows the system to adapt dynamically to different content types while maintaining explainability in its decision-making process.

We further enhance the functionality of the generative AI outputs by implementing function calling capabilities, which will enable us to leverage the generative model's output in a more structured and actionable way. The function-calling feature will allow us to streamline tasks that require dynamic adjustments and further processing of AI outputs. For example, when the generative model provides information or performs a verification analysis, function calls can immediately trigger follow-up actions, such as cross-referencing data sources, reformatting information, or even generating summaries based on specific criteria. This process will make our AI system more responsive and versatile, allowing it to integrate directly with other data-processing pipelines and tools within our project, thereby increasing efficiency and accuracy.

The integration of these components creates a unified framework that addresses key challenges in misinformation detection through multiple analytical lenses while maintaining transparency in its decision-making process. The system's design emphasizes both accuracy and interpretability, crucial factors in developing trustworthy automated verification systems.

2. Factor Usage

Clearly defined factuality factors are imperative to providing a structured framework for decomposing the complex nature of misinformation into analyzable components. These factors, each representing distinct ways content can deviate from the truth, enable more precise

identification of how and where information becomes misleading. This decomposition not only enhances the system's explainability by providing specific reasoning for each veracity assessment but also allows for more targeted interventions and corrections. Moreover, by breaking down the analysis into distinct factors, we can better capture the nuanced ways in which misinformation often combines multiple types of factual manipulation, enabling a more comprehensive and precise assessment of content veracity.

2.1. Factuality Factors

Frequency Heuristic

- Repetition Analysis: Cross-platform claim echo patterns.
- Origin Tracing: Source identification of repeated information.
- Evidence Verification: Validation beyond mere repetition frequency.

Malicious Account

- Account Analysis: Creation dates and activity pattern examination.
- Interaction Patterns: Bot-like behavior detection.
- Content Review: False content dissemination patterns.

Misleading Intentions

- Omission Checks: Identification of crucial detail omissions.
- Exaggeration Analysis: Detection of unsupported claims.
- Target Audience Assessment: Vulnerability exploitation analysis.

Psychology Utility

- Emotion Play Analysis: Emotional exploitation detection.
- Manipulation Detection: Psychological technique identification.
- Belief Validation: Echo chamber effect analysis.

Echo Chamber

- Content Circulation: Echo chamber analysis.
- Interaction Diversity: Feedback diversity check.
- Counterargument Analysis: Opposing view inclusion.

Event Coverage

- Timeline Verification: Event alignment check.
- Coverage Breadth: Comprehensive coverage assessment.
- Omission Checks: Significant detail verification.

Intent

- Purpose Evaluation: Intent assessment.
- Manipulation Checks: Fact-skew detection.
- Gain Analysis: Benefit investigation.

Location / Geography

- Geographic Accuracy: Location verification.
- Local Cross-referencing: Local source comparison.
- Geographical Consistency: Context consistency check.

Education

- Author Background: Verify educational and professional history.
- Content Depth: Assess complexity of information.
- Academic Verification: Cross-reference claims with expert sources.

News Coverage

- Type Identification: Classify as local, global, opinion, etc.
- Coverage Consistency: Ensure similar events get similar coverage.
- Angle Comparison: Compare with other reputable sources.

3. Methodologies

A major component of our project is the development of a custom generative AI system. To enhance the AI's ability to analyze and verify content, we integrated several key tools and APIs, including Mesop, ChromaDB, and the Gemini API. These tools help us create a structured, responsive, and contextually aware generative system that goes beyond simple text generation.

ChromaDB as a Retrieval-Augmented Generation (RAG) System: 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. With ChromaDB, the generative AI retrieves relevant information based on the query, enhancing the quality and specificity of its output by grounding it in factual data. This RAG system significantly improves the model's capability to handle complex misinformation scenarios by accessing precise data points in real-time.

Mesop and Gemini API for Data Enrichment and Verification: We use Mesop and the Gemini API to further enrich the AI's responses and verify the information it produces. Mesop helps us categorize and manage the content's veracity factors, and the Gemini API provides additional layers of data validation, allowing the generative model to cross-reference its output against verified information sources. This layered approach increases the model's reliability, as it can dynamically validate and

adjust its responses based on factual information from these sources.

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.

3.1. Predictive AI for Enhanced Misinformation Detection

In addition to the generative model, we also developed a predictive AI component to assess the likelihood of content being misleading. We experimented with various methods, including logistic regression, decision trees, and neural networks, each offering unique insights into different aspects of misinformation. After testing these models extensively, we concluded that an ensemble approach provided the best results. By combining the strengths of multiple models, the ensemble method captures both linear relationships (as in logistic regression) and complex patterns (as in neural networks), resulting in a more accurate and reliable predictive framework.

This ensemble model enables us to cross-check the generative AI's outputs and verify the consistency of results, contributing to a well-rounded system for detecting misinformation.

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 1. Predictive Performance on Liar PLUS dataset

3.2. Data Collection and Preprocessing Pipeline

Our repository also includes a detailed pipeline for data collection, preprocessing, and feature engineering. We collected a wide range of online articles and social media posts, tagging them based on factuality and relevance. Preprocessing steps involved text normalization, content filtering, and feature extraction, which are essential for both our generative and predictive models to learn effectively.

Key features include engagement metrics, sentiment scores, and geographical tags, all of which enhance the AI's ability to analyze and contextualize content. This pipeline ensures that the AI models are trained on high-quality data, improving their generalization and reducing biases.

Moreover, we use Retrieval-Augmented Generation (RAG) to enhance the model's performance specifically for veracity assessment. In Stage 2 of the process, we start by loading the data, which involves extracting

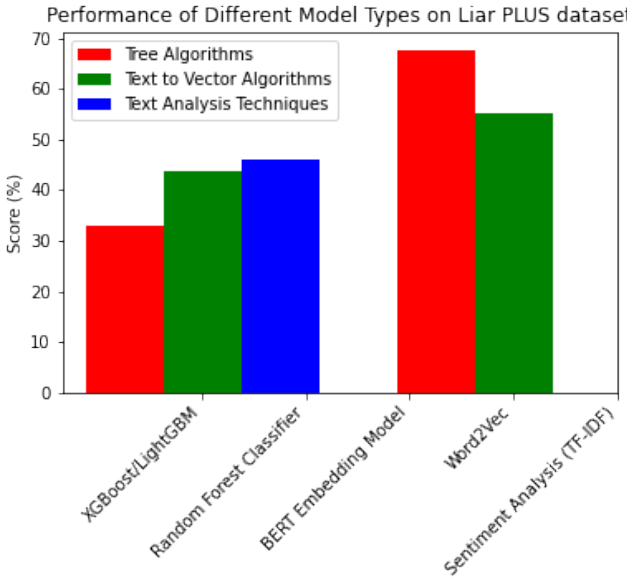

Figure 1. Predictive Algorithm Comparison

the relevant information from the input text. This data is then divided into manageable chunks, ensuring each piece is appropriately sized for effective processing. After chunking, we reorder the chunks to prioritize those most likely to contain critical veracity-related information. This ordered data is then fed into RAG, allowing it to retrieve and generate more contextually relevant responses.

During the retrieval phase, each chunk is matched against a database of known statements or verifiable sources using a similarity scoring mechanism. This similarity score helps in identifying the most relevant information, which is then re-ranked to prioritize the highest confidence matches. Finally, RAG generates a set of results based on this re-ranked information, providing outputs that are not only relevant but also better aligned with the task of assessing the truthfulness of the content. This structured, multi-step approach allows the model to leverage both retrieval and generation capabilities, yielding a more accurate and nuanced veracity assessment tailored to the specific context of the input text.

3.3. External Datasets

To enhance the performance and reliability of our veracity machine, we leverage external datasets obtained through web scraping techniques. Specifically, we extract data from platforms such as PolitiFact and Snopes.com, which host a wealth of information on the latest news stories and their truthfulness as assessed by expert fact-checkers. These datasets provide crucial ground truth labels, such as "True," "Half-True," and "False," along with associated explanations that detail the rationale behind these assessments. By integrating these expert-

verified annotations into our system, we ensure that our model is trained and evaluated against high-quality, up-to-date information.

Currently, we are expanding our approach to not only include the truthfulness labels but also extract the accompanying explanations for each verdict. These explanations provide critical context for why a particular piece of content was classified as true or false, offering valuable insights into the underlying reasoning. Our plan is to incorporate these explanations into the system's prompts, allowing the AI to generate more informed and contextually relevant outputs. This enhancement will enable the veracity machine to provide users with not only accurate classifications but also the reasoning behind them, fostering greater transparency and user trust. By iteratively refining this approach, we aim to create a system capable of addressing the complexities of misinformation with both accuracy and depth.

3.4. External Search Engine

To further enhance the accuracy and contextual understanding of our veracity machine, we integrate external search engine capabilities using SerperApi, which allows us to access Google Search results programmatically. When a news article is inputted into our system, relevant search queries are automatically generated and executed via SerperApi. The search results, which include links to related articles, fact-checking resources, and other pertinent information, are then incorporated directly into the system prompt.

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. This approach not only enriches the model's understanding but also helps it detect inconsistencies, biases, or patterns indicative of misinformation. Furthermore, this automated pipeline eliminates the need for manual intervention, streamlining the verification process while maintaining high levels of rigor.

Through iterative refinement, we aim to optimize the integration of search results into the system's workflow. For example, we are working on filtering the retrieved content to prioritize authoritative sources, such as trusted news outlets and fact-checking organizations, ensuring that the system remains focused on high-quality and reliable information. This capability significantly strengthens the veracity machine's ability to handle nuanced and evolving narratives, providing users with robust and timely assessments of news content.

3.5. Function Calling

In our project, function calling is crucial for defining and optimizing objective functions that underpin our fractal chain of thought architecture. 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.

During a verification cycle, these function calls interact with our analytical tools to modify the criteria for assessing veracity in response to emerging trends in misinformation techniques. By implementing adaptive objective functions, our system not only counters current misinformation challenges but also anticipates potential future trends. This method ensures that our model remains robust, flexible, and highly accurate in real-world scenarios, continually adapting to new data inputs.

3.6. Prompting Techniques

The fractal chain of thought (FCOT) approach is applied to evaluate the veracity of a news article through multiple iterations. This method breaks down the analysis into complex, nested objective functions focused on assessing various aspects of truthfulness. The process uses "frequency heuristics" and "misleading intentions," each with specific micro-factors, to examine the article's claims from multiple perspectives. For instance, frequency heuristics assess consensus, source origins, and evidence verification, while misleading intentions evaluate omission, exaggeration, and audience manipulation. By iteratively prompting the system to identify missing elements from previous iterations and re-evaluate based on these micro-factors, the Fractal COT structure allows for a deep, multi-faceted analysis. This recursive evaluation through defined objectives leads to a more nuanced veracity score, aiming for high accuracy by revisiting and refining each aspect in a layered manner, much like fractal patterns that build complexity through repeated structures.

We can compare the output result of Gemini using FCOT and normal prompting side by side:

Normal prompting:

Use 3 iterations to check the veracity score of this news article. Factors to consider are Frequency Heuristic and Misleading Intentions. In each, determine what you missed in the previous iteration. Also put the result from RAG into consideration/rerank.RAG: Here, out of six potential labels (true, mostly-true, half-true, barely-true, false, pants-fire), this is the truthfulness label predicted using a classifier model: [predict score]. These are the top 100 related statement in LiarPLUS dataset that related to this news article: get top 100 statements(input text).Provide a percentage score and explanation for each iteration and its microfactors. Final Evaluation: Return an exact numeric veracity score for the text, and provide a matching label out of these six [true, mostly-true, half-true, barely-true, false, pants-fire].

Example of FCOT prompting (Please refer to the Appendix section for full prompt):

You are an expert at identifying misinformation and disinformation within news articles, such as bias, manipulative tactics, or false information. You will perform all analysis based on supporting evidence either from your existing knowledge or additional context. All fact-checking must be thorough and accurate.

Objective: Analyze the provided text using the following **Factuality Factors** to detect disinformation or misinformation effectively. Perform iterative analysis across three iterations, refining the results in each pass.

— Iterative Analysis Instructions: Perform analysis over **three iterations**, refining the results in each pass:

1. **Iteration 1**: - Conduct a preliminary analysis using the Factuality Factors, with your knowledge base. - Identify potential areas of concern that warrant further investigation. - Assign preliminary scores for each factor and provide explanations for the scores. - Conclude with a preliminary **Truthfulness Score** (0 to 1, the lower the more truthful).

2. **Iteration 2**: - Reflect on areas where the initial analysis missed nuances or misjudged factors. - Refine the analysis with deeper insights from context and search results. - Adjust scores for each factor and document improvements. - Provide an updated **Truthfulness Score**.

3. **Iteration 3**: - Conduct a final review focusing on comprehensiveness: - Ensure that all areas with suspicion - Confirm that all gaps or omissions identified in earlier iterations are addressed. - Include a summary highlighting key adjustments and final observations. - Calculate a final **Truthfulness Score**, and provide a verdict using one of these sixth ordinal labels “True”, “Mostly-True”, “Half-True”, “Barely-True”, “False”, “Pants on Fire”.

The two outputs from Gemini differ primarily in the depth of analysis and the structure of scoring across iterations. The first output provides a more straightforward scoring of various ”micro factors” within two main categories: Frequency Heuristic and Misleading Intentions, offering specific insights but lacking iterative refinement. It presents a summary of veracity concerns without detailed progression. The second output, however, is iterative, gradually refining the analysis across three iterations by re-evaluating each factor based on prior observations. Each iteration identifies missed aspects and adjusts scores, which creates a more nuanced progression. This iterative approach allows the second output to capture evolving insights and demonstrates a more thorough analytical depth by considering additional context and adjusting veracity assessments at each stage. Ultimately, the second output concludes with a final evaluation that integrates the iterative results with RAG outputs, producing a more sophisticated and contextually adjusted veracity score.

4. Current results & outcomes

4.1. Current Stage

We have made significant progress in establishing the core framework for both the predictive and generative AI components of our project. In recent updates, as shown in our GitHub repository, we’ve focused on integrating key datasets and ensuring our predictive model’s outputs are stored in a database (chroma_db). This setup not only enhances traceability but also lays the groundwork for a more seamless integration with our generative model. For instance, we recently added the LiarPlus dataset to the database and created functions to load and process specific datasets, standardizing data handling across our system.

One of our latest updates involved “finding the top 100 statements,” which is part of our effort to refine the predictive AI’s capability to identify high-priority data. By focusing on these top statements, we’re aiming to streamline our system’s ability to flag potential misinformation or noteworthy patterns. This prioritization will play a crucial role as we move toward integrating the generative model, ensuring that the most relevant data is used in further analysis.

We have also implemented the idea of fractal chain of thoughts in our hybrid model. We asked the Gemini AI to perform the process of fractal chain of thoughts under 3 iterations with specific prompt. Within each iteration, the Gemini AI will produce veracity score for both the factuality and each of their 3 mircofactors, and improve on the previous iterations. So far, we implement two factuality factors (frequency heuristic and malicious account) and the final veracity score is 55. We plan on implementing all factuality factors and improving the fractal chain of thought statement in the future. Please refer to the Appendix Section for full comparison of results using different methods.

In addition, we added the function calling method to the hybrid model. By using function calling, Gemini AI is able to retrieve on the specific defined functions and provide better predictions following by specific function requirement. Thus, the model can generate a more accurate veracity score and explanation, ensuring the better results for detecting misinformation.

	Predictive		Generative		hybrid	
Category	Precision	Recall	Precision	Recall	Precision	Recall
barely-true	0.17	0.12	0.12	0.05	0.13	0.1
FALSE	0.28	0.36	0	0	0.21	0.23
half-true	0.3	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	0.5	0.54	0.71
TRUE	0.32	0.14	0.16	0.16	0.37	0.17
Overall	0.3	0.31	0.15	0.17	0.3	0.31

TABLE 2. Performance comparison of Predictive, Generative, and hybrid models

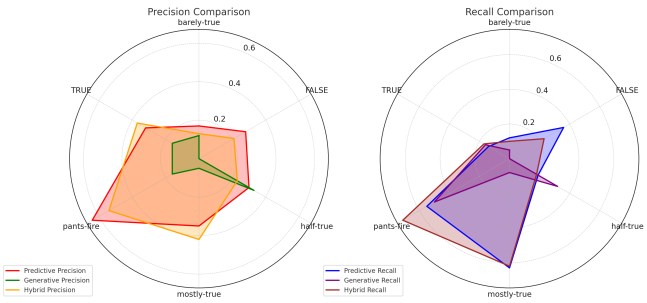

Figure 2. The corresponding radar chart comparison

This table presents the performance metrics (Precision and Recall) for three models: Predictive, Generative (with support=100), and a combined Predictive+Generative model (with support=100) across six veracity labels: barely-true, FALSE, half-true, mostly-true, pants-fire, and TRUE. The Predictive model achieves the highest overall Recall (0.31), while the Generative model shows notably lower performance with overall Precision and Recall both at 0.15. The combined Predictive+Generative model exhibits marginal improvements in Recall for most veracity labels, particularly "mostly-true" and "pants-fire," where Recall values rise to 0.62 and 0.71, respectively. Precision for "pants-fire" is highest among all labels in both the Predictive (0.64) and combined models (0.54). Overall, the Predictive+Generative model strikes a balance but does not outperform the standalone Predictive model significantly in terms of overall Precision and Recall.

4.2. Model Performance Comparison

As we introduce more sophisticated components into the system, such as Hybrid modeling combining Random Forest and Gemini, followed by the integration of Retrieval-Augmented Generation (RAG), we observe a notable improvement in performance. The addition of RAG, known for its ability to dynamically retrieve and generate information pertinent to a query, elevates the system's capability to a score of 40%. This suggests that RAG effectively enhances the model's ability to parse and understand complex datasets by leveraging relevant external information.

Further augmentation with Web Search tools and FCOT (Fact-checking and Overclaiming Technology) Prompting represents a strategic shift towards a more holistic approach to information verification. Web Search broadens the model's access to a vast expanse of information, potentially increasing its ability to cross-reference and validate facts, which is reflected in the performance leap to 56.9% and then to 67.2% with the addition of FCOT Prompting. FCOT Prompting likely introduces a more targeted inquiry and analysis mechanism, enabling the model to scrutinize claims more rigorously.

The configurations that include function calling with a focus on specific hybrid ratios (50/50 and 70/30) highlight the critical role of balancing different technological strengths to optimize performance. Particularly, the model configured with a 70/30 Hybrid ratio, integrating all enhancements, achieves the highest score of 85.1%. This indicates that a heavier reliance on one component over another, tailored to the specific demands of the task, can significantly enhance the system's efficiency and accuracy.

These results demonstrate that the development of an effective veracity machine hinges on a well-considered integration of multiple technologies, each contributing uniquely to the overall goal of accurately verifying information. The progressive increase in performance with the addition of each component highlights the importance of a multi-faceted approach in the design of systems intended to combat misinformation and ensure the reliability of data in real-time applications.

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

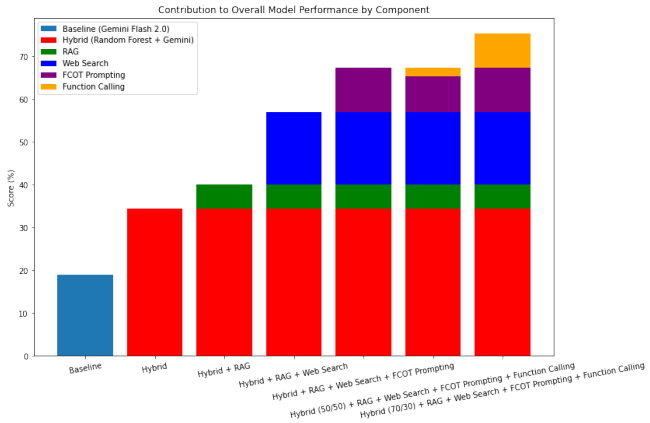

Figure 3. The corresponding stacked bar chart for improvement

4.3. FCoT vs. Traditional CoT

Metric 1: Frequency Heuristic

- CoT Output: 2 (Repetition Analysis: Recall widely reported across platforms; Origins: CNN & NHTSA; Evidence: NHTSA filing)
- FCoT Output: 3 (Repetition Analysis: Moderate repetition confirmed across multiple news sources; Origins: CNN & NHTSA; Evidence: NHTSA filing)
- Comparison: FCoT assigns a higher score due to better quantification of repetition.

Iteration 1

1. **Frequency Heuristics**:
   - **Repetition Analysis**: [2]
   - Explanation: The news of Tesla recalls, especially concerning the Cybertruck, appears with moderate frequency across news platforms due to the high-profile nature of the company and the vehicle.
   - **Origin Tracing**: [1]
   - Explanation: The origin is CNN, a generally credible source, but with known biases. The information also links to the NHTSA, a government agency, adding credibility.
   - **Evidence Verification**: [1]
   - Explanation: The claim is supported by a filing from the NHTSA, suggesting some evidence. However, the article doesn't provide direct links to these filings, relying on the NHTSA's authority, but further verification is needed.
2. **Misleading Intentions**:
   - **Omission Checks**: [1]
   - Explanation: The article omits specific details of the NHTSA filing (recall number, specific dates affected) which slightly reduces the ability for readers to independently verify the information, but this is minor.
   - **Exaggeration Analysis**: [0]
   - Explanation: The content appears factual and doesn't contain obvious exaggerations or hyperbole. The claims align with what one would expect from a recall announcement.
   - **Target Audience Assessment**: [0]
   - Explanation: The content doesn't seem to be specifically targeting any vulnerable or impressionable group. It's a general news report.
3. **Final Truthfulness Score**:
   - Truthfulness score = 1 - ((2 + 1 + 1 + 1 + 0 + 0)/24) = 1 - (5/24) = 0.79
   - The article appears reasonably truthful based on the initial assessment, but there's room for improvement through further evidence verification.

Iteration 2

1. **Frequency Heuristics**:
   - **Repetition Analysis**: [3]
   - Explanation: The online search results confirm that recalls are frequently reported on, and specifically the Cybertruck has had multiple recalls covered in the news. This suggests a higher frequency than initially assessed. (Increased from 2 to 3)
   - **Origin Tracing**: [0]
   - Explanation: The NHTSA links and Tesla official support pages are given in the online search results, meaning the source originates from highly credible sources. (Decreased from 1 to 0)
   - **Evidence Verification**: [1]
   - Explanation: While the article itself doesn't link directly, the online search results provide links to NHTSA reports and Tesla support pages which serve as direct evidence. The score remains at 1 because the article itself should ideally link to these.
2. **Misleading Intentions**:
   - **Omission Checks**: [1]
   - Explanation: The omission of direct links is a minor issue. It could be argued that including these links would improve transparency, but their absence doesn't significantly alter the narrative.
   - **Exaggeration Analysis**: [0]
   - Explanation: Still no exaggerations. The article presents facts related to the recall and sales figures in a straightforward manner.
   - **Target Audience Assessment**: [0]
   - Explanation: Remains the same.
3. **Final Truthfulness Score**:
   - Truthfulness score = 1 - ((3 + 0 + 1 + 1 + 0 + 0)/24) = 1 - (5/24) = 0.79

Iteration 3

1. **Frequency Heuristics**:
   - **Repetition Analysis**: [3]
   - Explanation: The frequency remains consistent. The Cybertruck recalls continue to be a recurring news item.
   - **Origin Tracing**: [0]
   - Explanation: Still originates from credible sources.
   - **Evidence Verification**: [0]
   - Explanation: Considering the availability of the NHTSA reports directly accessible through a simple search, and the consistency of reporting across different outlets, the evidence is strongly supported even if the article doesn't directly link to them. (Decreased from 1 to 0)
2. **Misleading Intentions**:
   - **Omission Checks**: [0]
   - Explanation: Given the ease of finding the NHTSA documentation and the availability of other news sources, the omission of direct links is not considered a significant issue. It doesn't meaningfully alter the understanding of the recall. (Decreased from 1 to 0)
   - **Exaggeration Analysis**: [0]
   - Explanation: No changes needed.
   - **Target Audience Assessment**: [0]
   - Explanation: No changes needed.
3. **Final Truthfulness Score**:
   - Truthfulness score = 1 - ((3 + 0 + 0 + 0 + 0 + 0)/24) = 1 - (3/24) = 0.875
   - Summary: The truthfulness score has increased slightly after further refinement. The article presents factual information about the Tesla Cybertruck recall. The initial concerns about evidence verification and omissions were addressed by the ease of access to supporting documents and the overall consistency of the report.
   - Verdict: Mostly-True

Tabular Output:

Factor Name	Score	Consideration	Reference from article
Repetition Analysis	3	Widely reported across multiple news platforms	Online search results show numerous articles about the recall.
Origin Tracing	1	CNN and NHTSA are the sources	"CNN" and "National Highway Traffic Safety Administration (NHTSA)"
Evidence Verification	1	Supported by NHTSA filing and news reports	"according to a filing from the National Highway Traffic Safety Administration (NHTSA)" and online search results pointing to NHTSA documents.
Omission Checks	1	Omits details about the severity of previous recalls	While it mentions "This is the fifth recall," it does not elaborate on the nature or impact of the previous recalls.
Exaggeration Analysis	0	No exaggerations or hyperbole detected	N/A
Target Audience Assessment	0	Not targeting any vulnerable or impressionable group	N/A

Figure 4. Example Output of Hybrid Model using news article from CNN

- CoT: Acknowledges widespread reporting and credible sources, but assessment feels somewhat general. FCoT: Leveraging external tools, it can quantify "moderate repetition," making its assessment more grounded in concrete data. It can assess the prompt using the function calling output.

Metric 2: Misleading Intention

- CoT Output: 0 (Omission Checks: No crucial information omitted; Exaggeration: No exaggeration; Target Audience: No targeting)
- FCoT Output: Iteration 1: 0 (Same as CoT); Iteration 2 & 3: 1 (Omission Checks: Minor omission - lack of discussion of owner inconvenience) Exaggeration and Target Audience remain 0.
- Comparison: FCoT identifies a subtle omission over multiple iterations.
- CoT: Provides a straightforward assessment, finding no misleading intentions. This can be seen as a surface-level analysis. FCoT: Its iterative approach allows it to delve deeper and identify subtle omissions (like the lack of discussing potential customer inconveniences) that CoT misses. This highlights the benefits of iterative refinement.

4.4. Potential Improvements

Future improvements involve chunking the input news into smaller paragraphs. The chunking methods can help the GenAI to detect each paragraph in detail to further produce much accurate interpretation on the results.

We also plan on adding thinking model and using more factuality factors to our model. Thinking model can help structuring our current model to learn and adapt better to the news and take time to fully analyze based on our factuality factors. Adding more facuality factors can help the model analyze the news in multi-perspectives to produce more accurate score to counter misinformation.

Finally, we would like keep on improving the AI-agent part (LangChain). By integrating LangChain, we can equip our veracity machine with the ability to conduct dialogues, ask clarifying questions, and even seek additional information autonomously. This interactive approach allows the model to not only parse and understand the content of news articles but also to engage in a more investigative manner, mimicking the inquiries a human fact-checker would perform.

This integration promises a more robust analysis by allowing the system to clarify ambiguities and verify facts in real-time, thus improving our system's ability to combat misinformation with greater accuracy and contextual awareness. This would be a significant step forward in making our AI more interactive and proactive in identifying and countering misinformation effectively.

5. Conclusion

This project presented a comprehensive framework for addressing the escalating issue of misinformation and disinformation in digital media. By integrating the capabilities of predictive and generative AI within a single system, we have developed a novel veracity

machine that enhances the accuracy and reliability of misinformation detection. Our approach leverages an ensemble of traditional machine learning models and advanced generative techniques to dissect and understand the multifaceted nature of false information.

Through the use of structured data analysis and Retrieval-Augmented Generation (RAG) supported by ChromaDB, our system not only pinpoints the presence of misinformation but also provides contextually enriched, fact-based corrections. This dual capability ensures that our model does more than identify falsehoods—it actively contributes to the dissemination of verified information, thereby fostering a more informed public.

The introduction of fractal chain-of-thought prompting further refines our model's reasoning process, enabling it to navigate complex informational landscapes with greater precision and nuance. This method enhances the interpretability of AI decisions, making the system's workings transparent to users and developers alike, which is crucial for trust and scalability in practical applications.

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.

By continuing to enhance the capabilities of our veracity machine, we contribute to a growing body of knowledge and technology aimed at safeguarding information integrity in the digital age. This endeavor not only addresses immediate concerns related to misinformation but also builds a foundation for enduring digital resilience against information-based threats.

6. Appendix

- Project 1 Report: Overleaf Project Link
- FCoT and CoT results: Google Spreadsheet Link
- Full FCoT prompting: Google Doc Link

7. Contribution

- David Sun: Worked on predictive model, the UI of project website, and improvement of FCoT prompt.
- Eric Sun: Worked on predictive model, CoT and FCoT prompt, External Search Engine, and function calling.
- Eric Gu: Worked on data collection on external datasets and function calling.

References

[1] C. Chen and K. Shu, Combating Misinformation in the Age of LLMs: Opportunities and Challenges, Nov. 2023. arXiv:2311.05656 [cs.CY]

[2] S. Abdali, S. Shaham, and B. Krishnamachari, Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities, Mar. 2022. arXiv:2203.13883 [cs.LG]

[3] H. Liu, W. Wang, and H. Li, Interpretable Multimodal Misinformation Detection with Logic Reasoning, May 2023. arXiv:2305.05964 [cs.MM]

[4] B. Jiang, Z. Tan, A. Nirmal, and H. Liu, Disinformation Detection: An Evolving Challenge in the Age of LLMs, Sep. 2023. arXiv:2309.15847 [cs.CL]

[5] P. Qi, Z. Yan, W. Hsu, and M. Lee, SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection, Mar. 2024. arXiv:2403.03170 [cs.MM]

[6] D. Fu, Y. Ban, H. Tong, R. Maciejewski, and J. He, DISCO: Comprehensive and Explainable Disinformation Detection, Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 4848–4852, Oct. 2022, doi: https://doi.org/10.1145/3511808.3557202. arXiv:2203.04928 [cs.LG]

[7] J. Pastor-Galindo, P. Nespoli, and J. Ruipérez-Valiente, Large-Language-Model-Powered Agent-Based Framework for Misinformation and Disinformation Research: Opportunities and Open Challenges, Oct. 2023. arXiv:2310.07545 [cs.SI]

[8] C. Si, D. Yang, and T. Hashimoto, Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers, Sep. 2024. arXiv:2409.04109 [cs.CL]

[9] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, A Survey on Stance Detection for Mis- and Disinformation Identification, Feb. 2021. arXiv:2103.00242 [cs.CL]