Towards Developing an Agent-Based Framework for Validating the Trustworthiness of Large Language Models

Johannes Bubeck, Janick Greinacher, Yannik Langer, Tobias Roth and Carsten Lanquillon[a]
Business Information Systems, Heilbronn University of Applied Sciences, Heilbronn, Germany

Keywords: Large Language Models, ChatGPT, Prompting, Trustworthiness, Validation, Verification.

Abstract: Large language models (LLMs) have revolutionized the field of generative artificial intelligence and strongly affect human-computer interaction based on natural language. Yet, it is difficult for users to understand how trustful LLM outputs are. Therefore, this paper develops an agent-based framework by exploring approaches, methods, and the integration of external data sources. The framework contributes to AI reasearch and usage by enabling future users to consider LLM outputs more efficiently and critically.

1 INTRODUCTION

Large Language Models (LLMs) have significantly transformed the fields of artificial intelligence (AI) and natural language processing (NLP), leading to the creation of advanced conversational models. With the increasing complexity and capabilities of these models, the necessity to ensure their outputs are trustworthy and verifiable becomes paramount (Ali et al., 2023). Trust is increasingly recognized as an essential element in user-AI interactions, especially as more opaque and complex machine learning models become widespread in AI (Jacovi et al., 2021).

Moreover, the field of LLMs is experiencing an exponential growth in the number of papers published each month (Zhao et al., 2023). As LLMs gain more prominence, the validation of their outputs for trustworthiness becomes increasingly crucial. This trend aligns with the strong focus on ethical considerations in AI and machine learning, urging the need to ensure trustworthiness in both current and future AI systems and applications (Future of Life Institute (FLI), 2021).

A viable solution to this challenge is the validation of LLM outputs. This validation process ensures that the models operate as intended, producing accurate and reliable results. Validation, in this context, involves verifying the trustworthiness of a model's outputs against established criteria or standards (Bowman and Dahl, 2021; Srivastava et al., 2022).

GPT-4 from OpenAI is one of the most advanced language models currently available. It vigorously exemplifies the critical need for validating LLM outputs. Although ChatGPT, with its web interface that facilitates easy access to different GPT versions, has shown remarkable ability in generating coherent and contextually appropriate responses to user queries, instances of the model delivering inconsistent outputs have also been reported (Jang and Lukasiewicz, 2023).

Hence, validating the outputs of LLMs is vital to ensure their reliability across various scenarios. This leads to our research question:

> How can we construct a framework that enhances LLM outputs by incorporating existing validation approaches and external data sources to increase their trustworthiness?

To advance our research, we have identified the following subsidiary questions:

1. What suitable validation approaches and external data sources already exist in the literature?
2. What are a reasonable architecture and process for developing such a framework?
3. How can a prototype be developed based on this framework?

The primary contribution of this paper is the development of an agent-based framework that augments the outputs of LLMs, providing users with feedback on the trustworthiness of these outputs. We have developed a prototype to test and evaluate our proposed framework. To address these research questions, we have created the following artifacts:

- a literature review identifying suitable validation approaches utilizing external data,

[a] https://orcid.org/0000-0002-9319-1437

Bubeck, J., Greinacher, J., Langer, Y., Roth, T. and Lanquillon, C.
Towards Developing an Agent-Based Framework for Validating the Trustworthiness of Large Language Models.
DOI: 10.5220/0012364000003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 527-534
ISBN: 978-989-758-680-4; ISSN: 2184-433X

- a multi-agent validation framework,
- a prototype implementation of the framework, and
- an evaluation of the prototype.

The structure of this paper is organized as follows: First, we discuss the scientific foundations and theoretical background. This is followed by a detailed presentation of our research methodology and our findings. Finally, we conclude with a discussion of our results and an exploration of potential future research avenues.

2 THEORETICAL BACKGROUND

2.1 Large Language Models

LLMs are advanced AI models with an extensive number of parameters. They are pre-trained on vast text corpora using self-supervised learning methods, and are often further refined with task-specific examples in supervised settings. Their development began to significantly advance around 2018, with both their parameter count and performance showing exponential growth since then (Birhane et al., 2023).

Current LLMs employ deep neural networks and leverage sophisticated machine learning techniques, such as the transformer model (Vaswani et al., 2017). They utilize substantial computational resources and large datasets, including internet content, enabling them to develop a comprehensive understanding of language and respond in contextually relevant and coherent ways (Teubner et al., 2023).

A major advantage of LLMs is their versatility in tasks like text composition, question answering, translation, chatbot operations, and program code generation. They are distinguished by their scalability and high parameter counts, which contribute to their cutting-edge performance in natural language processing. Notably, LLMs can initially learn from unlabeled data, a self-supervised learning approach that yields impressive outcomes and broadens their applicability without extensive supervised training (Vaswani et al., 2017; Huang et al., 2022).

However, there are several concerns associated with LLMs. These include challenges in capturing scientific interpretation and meaning, potential neglect of value judgments in scientific texts, risks of producing inaccurate content, so-called hallucinations, and the potential erosion of trust in the peer review process. These issues underscore the need for cautious and critical assessment, as well as further research to ensure responsible and appropriate use of LLMs in scientific settings (Birhane et al., 2023).

2.2 Trust in Artificial Intelligence

Trust is a fundamental factor in the adoption and effective utilization of AI technologies. This is particularly true for the complex models used in natural language processing, where high levels of trust are imperative. Trust enables researchers, developers, and users to depend on the accuracy and reliability of AI-generated results. It lays the groundwork for applying AI in diverse areas, including machine translation, text generation, and information retrieval. To establish trust, an AI system's transparency, interpretability, and accountability need to be thoroughly addressed. This is essential to ensure that an AI system is not only reliable but also ethically responsible (Toreini et al., 2020).

However, the task of ensuring AI systems' trustworthiness is fraught with challenges, primarily due to the inherent *black box* nature of these systems. In particular, LLMs with their complex deep neural network architectures and extensive parameter spaces are trained on massive datasets. The opacity of these deep learning systems presents a significant challenge, as their decision-making processes typically remain elusive. The lack of clarity and understanding of these processes casts doubt on the extent to which these systems can be trusted. As such, transparency is paramount in building trust in AI, ensuring that users can have confidence in their outcomes (von Eschenbach, 2021).

3 RESEARCH METHODOLOGY

For the scientific design of our research process, we adhere to the established methodology outlined in (Peffers et al., 2007). Our research process is segmented into five distinct steps, as depicted in Figure 1 and described in greater detail below.

Our literature review methodology is guided by (Webster and Watson, 2002) and (vom Brocke et al., 2009). We conducted title searches using the terms ["fake checking" AND ("methods" OR "approaches")] and ["fact checking" AND ("methods" OR "approaches")], with no date restrictions, across several literature databases including IEEE Xplore Digital Library, ScienceDirect, SpringerLink, and Emerald Insight. In addition, we conduct forward and backward searches (Webster and Watson, 2002).

In the second step, relevant papers are read thoroughly and, also, implemented. For the coding process of the approaches, the procedure of qualitative content analysis as proposed by (Mayring and Fenzl, 2010) is performed. All publications are analyzed to

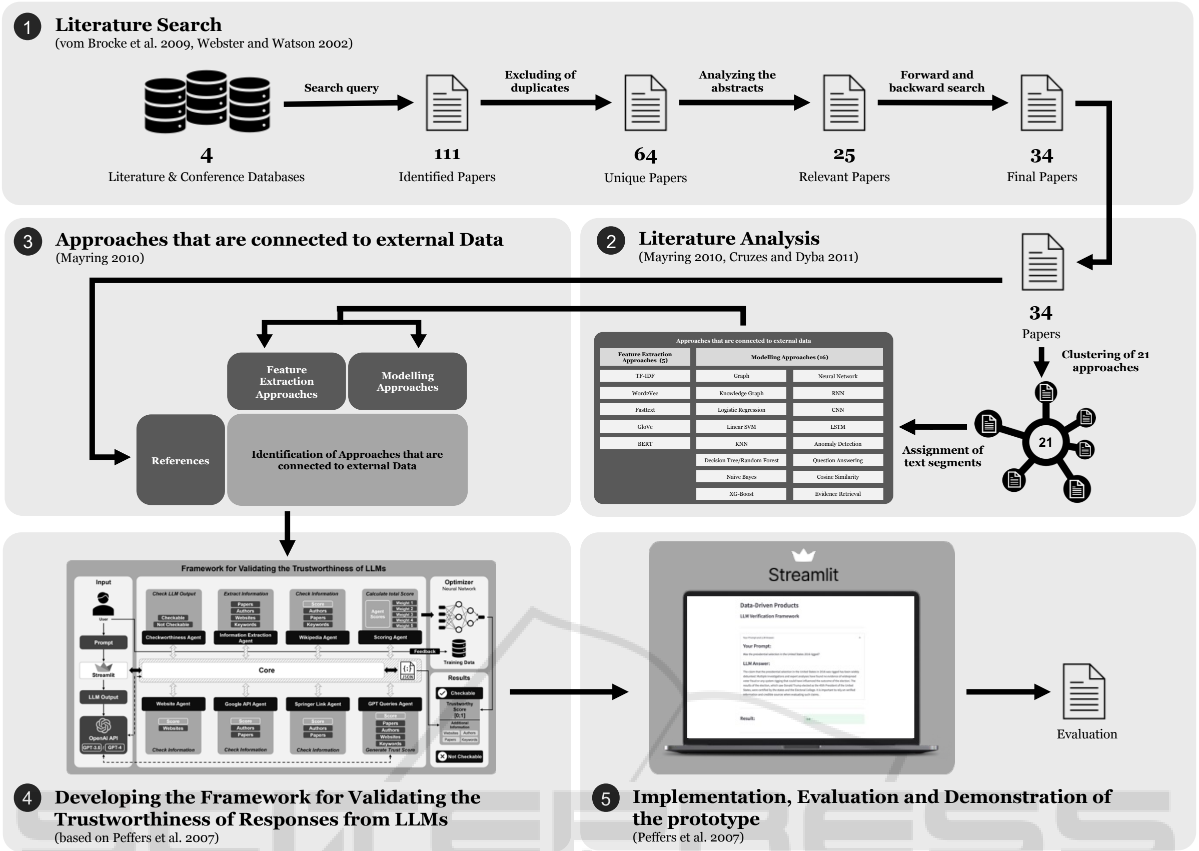


Figure 1: Overview of the research process (Peffers et al., 2007).

identify approaches and external data sources. Only text segments describing approaches connected to external data are considered here. Based on this, we identify several text segments according to the thematic synthesis process by (Cruzes and Dyba, 2011). These segments are condensed into various characteristics and assigned to high-level categories (Mayring and Fenzl, 2010). Subsequently, in the third step, we summarize the segments and the associated publications in a concept matrix (Webster and Watson, 2002).

The fourth step is dedicated to conceptualizing the framework. This entails analyzing the identified approaches related to external data sources and integrating them within the framework. A critical aspect of this conceptual design is the recognition that not all identified approaches could be considered in the initial implementation of the framework. However, the agent-based structure that we propose will offer the flexibility to easily incorporate additional approaches in future iterations.

In the final step, the design and development and, also, the demonstration and evaluation phases are completed according to (Peffers et al., 2007). The prototype is being implemented, evaluated, and demonstrated based on our proposed framework. Particular emphasis is placed on the structure of the agents, ensuring that they align well with the overarching framework and meet the objectives set out in our research.

4 FINDINGS

This section delves into the detailed presentation and discussion of the results and findings. It encompasses the literature review, the conceptual design of the framework, and the development of the prototype.

4.1 Literature Review

The literature search process resulted in 111 publications. After a thorough review of the abstracts and keywords, we narrowed down to 25 publications that were directly relevant to our research focus. The forward and backward searches increased the final number of pertinent publications to 34.

We identified two main categories: *feature extraction* and *modeling approaches*. These categories form the foundation for conceptualizing and developing both the framework and the prototype. As depicted in Figure 1 (step 2), the category *feature extraction* comprises five approaches, while *modeling approaches* includes 16.

We conducted a comprehensive analysis and evaluation of these approaches to assess their suitability. Some approaches are incompatible with our current research focus, which is primarily focused on the conceptualization of a prototypical framework. Thus, we carefully selected suitable approaches and integrated them into the initial iteration of our project. Nonetheless, the knowledge gained from the literature review is expected to be extremely valuable for future research and the continued development of the framework in its subsequent iterations.

4.2 Framework

The development of our framework is based on the results and analyses of approaches identified in the literature review. We opted for an agent-based framework as the foundational conceptual model because it allows seamless integration of the approaches we identified without the need for complex adaptations. This ensures the framework's modularity and scalability. We consider agents as independent entities, each with specific tasks, capable of interacting within the framework. The analysis of the trustworthiness of LLM outputs is facilitated by individual agents, each calculating scores for different tasks. A neural network, referred to as the *scoring agent*, is then used to compute a weighted overall score.

For the frontend, we selected Streamlit due to its ability to support the rapid development of visually appealing prototypes. Figure 2 illustrates our framework, which will be discussed in more detail in the subsequent sections.

The *OpenAI API* component serves as a gateway to the OpenAI API, enabling interaction with GPT-3.5-Turbo or GPT-4 for text generation. This interface facilitates sending requests to the OpenAI API and receiving the corresponding text responses. As a part of our research, we approached OpenAI to share our research methodology and were subsequently granted access to GPT-4, allowing us to test our framework with its more sophisticated outputs.

The *checkworthiness agent* assesses the reliability of statements utilizing the OpenAI API to process any request. Its *run* method analyzes the model's response and produces a list indicating a statement's verifiability along with detailed explanations.

Core is the central component of the LLM validation framework. It orchestrates various agents, such as the checkworthiness agent, information extraction agent, website agent, Wikipedia agent, Google API agent, GPT queries agent, and scoring agent. Triggered from the Streamlit frontend after confirming a statement's trustworthiness, the core's *main* function orchestrates the sequence flow. This enables the integration of data from different agents and acts as the central control point of the framework.

In the *information extraction agent*, an interface connects to the OpenAI API to conduct specific searches for papers, authors, websites, and keywords used in the model's outputs. This agent processes this information with comprehensive error handling to address potential inaccuracies or incorrect feedback.

The *Wikipedia agent* extracts information using keywords from the information extraction agent as search queries. It scans the initial articles using these keywords to identify authors' names and paper titles. It then compares these details with known authors and titles. After this comparison, a scoring mechanism is applied to measure the level of similarity or relevance between them.

The *website agent* scores a given LLM output based on matching websites and a thorough analysis based on their content and status. The agent sends requests to individual websites and analyzes the resulting HTTP status codes as they indicate whether a website is functioning correctly, undergoing redirection, offering informative content, or facing errors. The final score is determined based on the number of matching websites found and their status categorization.

The *Google API agent* leverages the Google Books API to review books and their respective authors. It examines authors and book titles by dispatching API requests, evaluating the outcomes, and compiles distinct lists of verified authors and books. Finally, the agent calculates a score based on the frequency of successful Google Books API queries.

The *SpringerLink agent* gathers information through an advanced search on the SpringerLink website, explicitly searching for certain authors and titles. The results it receives are then compared with the data extracted by the information extraction agent, which includes details of authors and papers. Similar to the process in the Wikipedia agent, a score is calculated for this agent based on the number of successful matches found between the authors and papers identified by the information extraction agent.

The *GPT queries agent* employs the OpenAI model to generate trust scores. This agent initializes the OpenAI object and uses a predefined text as a template for evaluating trust. Unlike other agents that derive their scores from external data, this agent calcu-

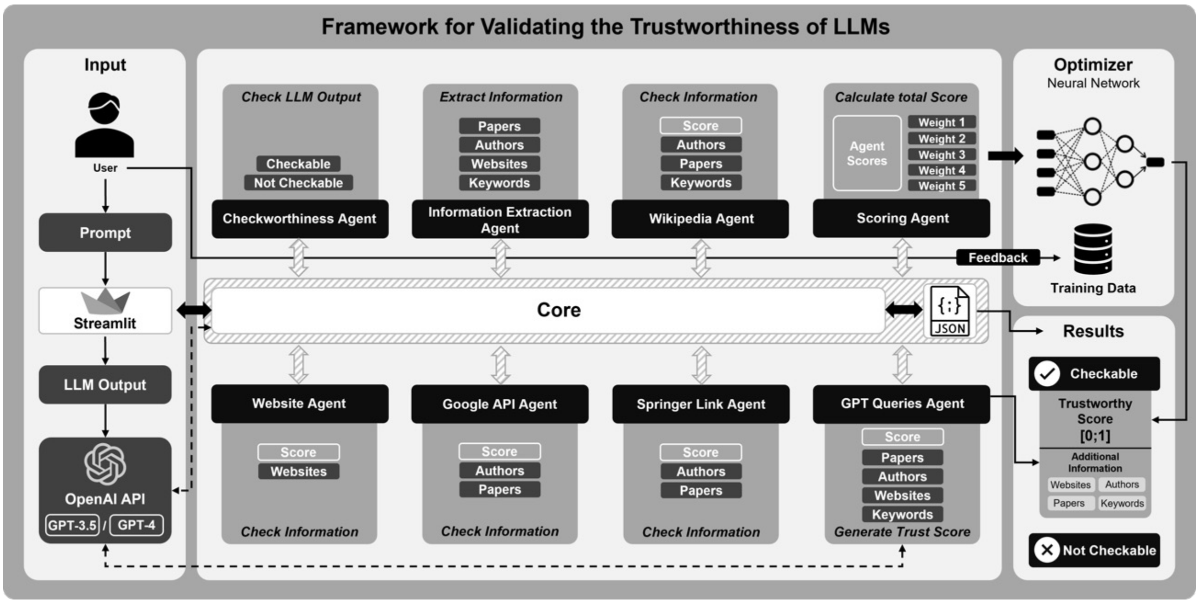


Figure 2: Overview of identified approaches.

lates the LLM's trustworthiness score internally. The LLM assesses this score using information extracted by the information extraction agent. Therefore, its inputs are solely its own previously generated outputs, without any external data.

The *scoring agent* and the *optimizer* component are responsible for calculating the overall trustworthiness score and optimizing it. To do so, the scoring agent activates a neural network module with the individual agents' scores. The overall trustworthiness score is computed as a weighted sum of the individual agent scores. A neural network optimizes the weights by predicting the specific weight for each agent's score based on the available data. Users can contribute feedback on the trustworthiness score's accuracy, which is then incorporated into the neural network's training data. This retraining occurs when the framework is initially launched.

Currently, the overall trustworthiness score includes assessments from five agents, each determining its score individually. These individual scores are based on how closely their findings align with the LLM's output, for which the trustworthiness score is being calculated. The scores range from 0 to 1, with 0.5 indicating a neutral position. Each score reflects the degree to which the LLM's output is considered trustworthy. If an agent is unable to gather useful information, it adapts by assigning a neutral score of 0.5. To maintain transparency and comparability, the individual agents' algorithms for calculating their scores always follow a consistent pattern based on the numbers of matches and their respective quality.

4.3 Implementation of the Prototype

The implementation of the prototype closely follows the conceptual framework's description. Figure 3 displays a section of our sequence diagram. Given the prototype's complexity, which includes a multitude of agents and their interactions, the sequence diagram is essential for structuring the prototype and the code flow and enhancing its comprehensibility.

With Python being the programming language of our choice, an object-oriented approach was adopted, and each agent is implemented as a distinct object or class. JSON was employed to consolidate the results and scores, with only the core module having administrative rights. Once the framework is initialized, the core module assumes full control over all other modules, orchestrating the individual results from the agents.

The validation process is initiated only if the LLM's statement is deemed verifiable by the checkworthiness agent. Access to the LLM is facilitated through the OpenAI API, where requests are sent to the designated chat completion endpoint. Authentication to the API is secured using organization details and API keys, which are also essential for API billing purposes. This combination of authentication methods is crucial to ensure that the model operates within the context of the initially provided prompt, which is key to our prompt design strategy. As mentioned in the previous section, we gained access to the GPT-4 model from OpenAI for this research project, following a request and a description of our research focus.

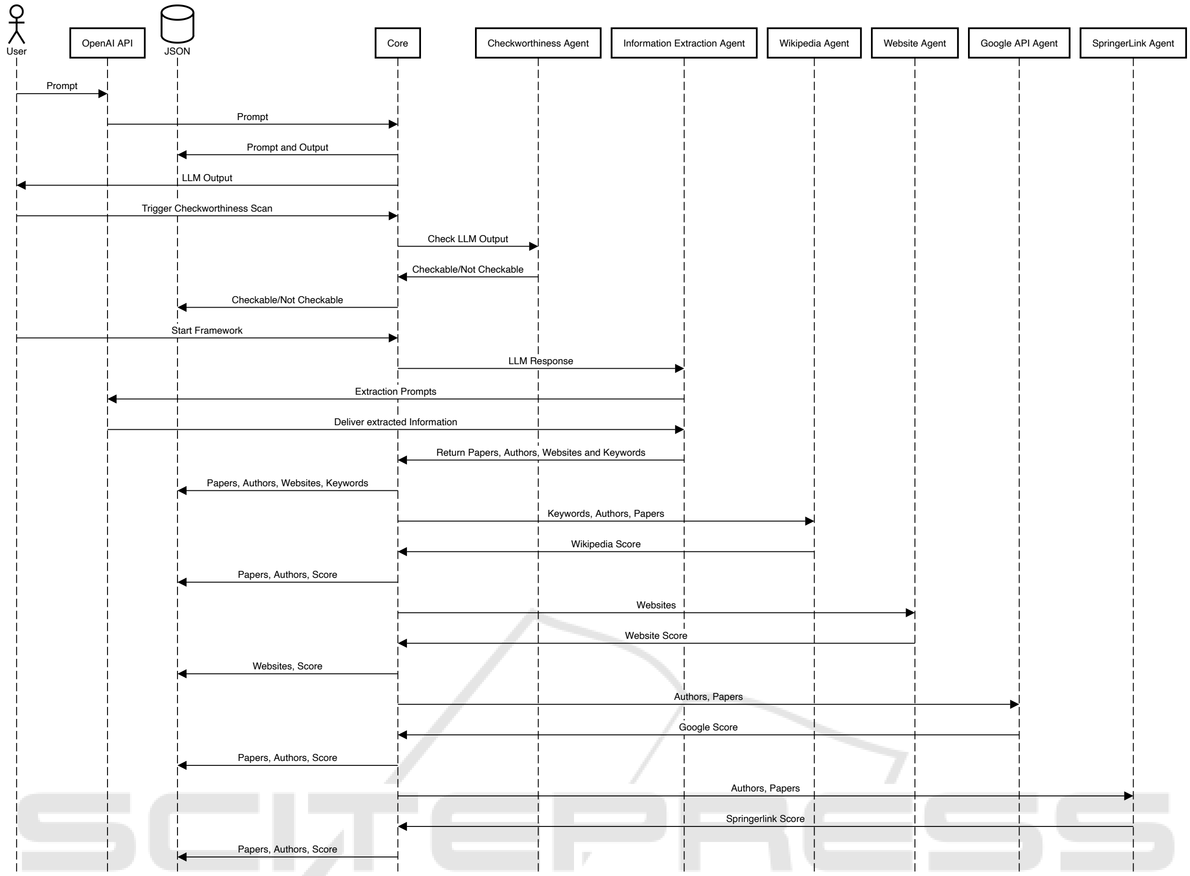


Figure 3: Part of the developed sequence diagram.

5 EVALUATION

This section focuses on evaluating the developed framework and its prototype implementation. It aims to determine the feasibility of successfully evaluating the trustworthiness of LLM outputs.

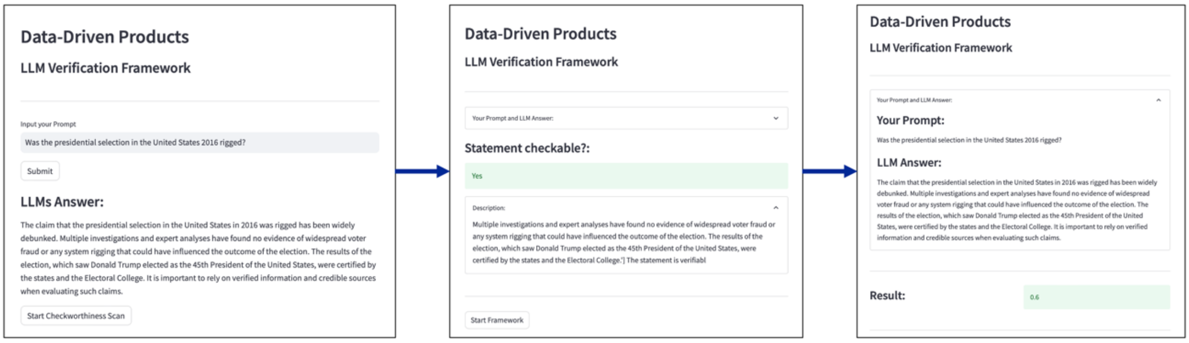
The user workflow and the prototype are depicted in Figure 4, using the example of querying whether the 2016 U.S. presidential election was rigged. The prototype, created in Streamlit, functions as a local web server. Users can input any text or question into the start screen as requests for a LLM. Once a prompt is submitted, it is sent to the OpenAI API, and the response from OpenAI's GPT-4 model is displayed. Initially, the user has the option to check if the output is worthy of further scrutiny. A reasoned response is then provided, indicating whether the output can be checked. If the LLM's output is deemed uncheckable, the user can initiate a new query. Otherwise, the user can activate the framework, which then starts the complex validation process in the background. This involves the individual agents' data processing and information analysis. Once all agents have completed their tasks and the overall trustworthiness score is calculated, it is displayed appropriately.

Besides the output results, users can view all the underlying information used in the agents' calculations, as illustrated in Figure 5. The detailed results are provided for questions such as: "Was the moon landing real?"

Access to detailed information about the output and the corresponding score enables users to not only receive the LLM's answer to their initial prompt, but also gain additional insights about it. This approach empowers users to delve into and scrutinize the relevant sources and authors associated with the information provided.

6 CONCLUSION

Our research methodology has been an effective foundation for developing an agent-based framework to validate the trustworthiness of LLMs. In this section, we summarize and critically analyze the results, while



Data-Driven Products

LLM Verification Framework

Input your Prompt

Was the presidential selection in the United States 2016 rigged?

LLMs Answer:

The claim that the presidential selection in the United States in 2016 was rigged has been widely debunked. Multiple investigations and expert analyses have found no evidence of widespread voter fraud or any system rigging that could have influenced the outcome of the election. The results of the election, which saw Donald Trump elected as the 45th President of the United States, were certified by the states and the Electoral College. It is important to rely on verified information and credible sources when evaluating such claims.

Data-Driven Products

LLM Verification Framework

Statement checkable?:

Yes

Description:

Multiple investigations and expert analyses have found no evidence of widespread voter fraud or any system rigging that could have influenced the outcome of the election. The results of the election, which saw Donald Trump elected as the 45th President of the United States, were certified by the states and the Electoral College.'] The statement is verifiabl

Data-Driven Products

LLM Verification Framework

Your Prompt:

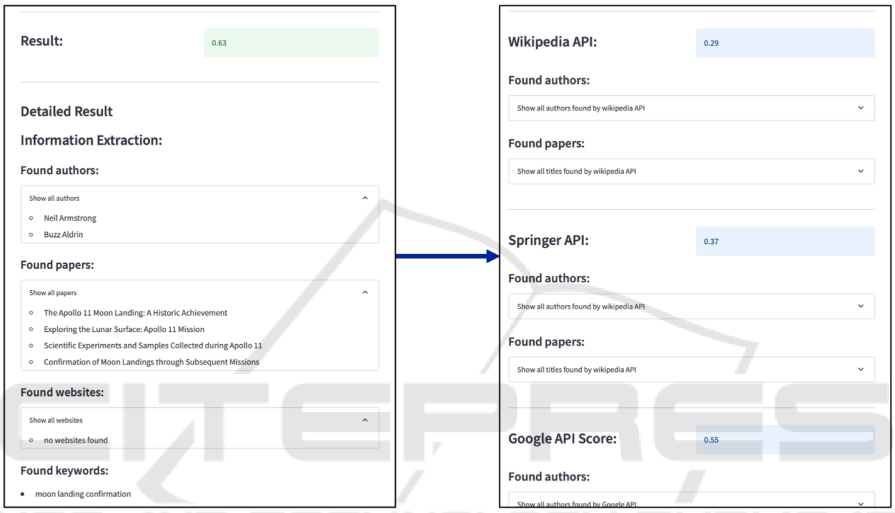
Was the presidential selection in the United States 2016 rigged?

LLM Answer:

The claim that the presidential selection in the United States in 2016 was rigged has been widely debunked. Multiple investigations and expert analyses have found no evidence of widespread voter fraud or any system rigging that could have influenced the outcome of the election. The results of the election, which saw Donald Trump elected as the 45th President of the United States, were certified by the states and the Electoral College. It is important to rely on verified information and credible sources when evaluating such claims.

Result: 0.6

Figure 4: Interaction process of the user with the framework.



Result: 0.63

Detailed Result

Information Extraction:

Found authors:

- Neil Armstrong
- Buzz Aldrin

Found papers:

- The Apollo 11 Moon Landing: A Historic Achievement
- Exploring the Lunar Surface: Apollo 11 Mission
- Scientific Experiments and Samples Collected during Apollo 11
- Confirmation of Moon Landings through Subsequent Missions

Found websites:

- no websites found

Found keywords:

- moon landing confirmation

Wikipedia API: 0.29

Found authors:

Found papers:

Springer API: 0.37

Found authors:

Found papers:

Google API Score: 0.55

Found authors:

Figure 5: Detailed results of the calculated trustworthiness score.

also identifying potential avenues for future research.

In our analysis of the literature, we identified 21 potential approaches for prototypically implementing an agent-based framework. By adopting the selected methodologies and integrating external data sources, our prototype framework computes a trustworthiness score. This score is calculated based on the varying weights assigned to the evaluations from individual agents. It serves as a measure for assessing the reliability of OpenAI's GPT-4 model in generating responses. The trustworthiness score, which ranges from 0 to 1, acts as a metric to determine the degree of trust that can be placed in the model's outputs. A score of 0 indicates a low level of trustworthiness, whereas a score of 1 represents an exceptionally high level of trustworthiness.

The agent-based framework we have conceptualized offers a flexible and scalable architecture for future development. Evaluating the system through our implemented prototype has yielded valuable insights into its functionality and usability. Since not all agent approaches identified in our literature review have been included in the prototype, there is room for integrating additional agents in future iterations.

As stated above, our research methodology was effective in developing the implementation of the agent-based framework. However, this work is not without its limitations. Additional approaches could have been explored for use within the framework to validate the trustworthiness of LLM outputs. While we have successfully incorporated various approaches in the form of agents, there is currently no scientifically robust justification for the selection of these approaches. Moreover, our framework is currently limited to six agents due to its prototype status. Therefore, future developments could involve expanding the framework by adding new agents, potentially developing and implementing a separate agent for each method identified in our literature analysis.

To the best of our knowledge, this paper presents a novel approach for validating the trustworthiness of LLMs. The findings here lay the groundwork for subsequent research. Future studies might focus on enhancing the proposed trustworthiness score calculation. For instance, our neural network could undergo further iterations to incorporate diverse data sources, thereby optimizing the weighting of agents in determining the score. Additionally, future work should involve reconfiguring the architecture of the prototype and subjecting it to empirical evaluation.

In the scope of our work, we have successfully created a prototypical agent-based framework for assessing the trustworthiness of LLMs. This prototype establishes a robust foundation and signals promising directions for future advancements. With this implementation, we have made a contribution in developing a validation framework for LLM outputs, marking a vital step towards its potential future application.

REFERENCES

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805.

Birhane, A., Kasirzadeh, A., Leslie, D., and Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280. Number: 5 Publisher: Nature Publishing Group.

Bowman, S. R. and Dahl, G. (2021). What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Cruzes, D. S. and Dyba, T. (2011). Recommended steps for thematic synthesis in software engineering. In *2011 international symposium on empirical software engineering and measurement*, pages 275–284. IEEE.

Future of Life Institute (FLI) (2021). The Artificial Intelligence Act.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. (2022). Large Language Models Can Self-Improve. arXiv:2210.11610 [cs].

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 624–635, New York, NY, USA. Association for Computing Machinery.

Jang, M. and Lukasiewicz, T. (2023). Consistency Analysis of ChatGPT. arXiv:2303.06273 [cs].

Mayring, P. and Fenzl, T. (2010). Qualitative inhaltsanalyse [qualitative content analysis]. *Qualitative Forschung Ein Handbuch (Qualitative Research: A Handbook)*, pages 468–475.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3):45–77. Publisher: Routledge _eprint: https://doi.org/10.2753/MIS0742-1222240302.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., and Hinz, O. (2023). Welcome to the Era of ChatGPT et al. *Business & Information Systems Engineering*, 65(2):95–101.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., and van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 272–283, New York, NY, USA. Association for Computing Machinery.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. arXiv:1706.03762 [cs].

vom Brocke, J., Simons, A., Niehaves, B., and Reimer, K. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Proess. In *European Conference on Information Systems (ECIS) 2009*.

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 34(4):1607–1622.

Webster, J. and Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2):xiii–xxiii.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023). A Survey of Large Language Models. arXiv:2303.18223 [cs].