A Comprehensive Review of Enhancing Open Foundation Models with NLP Techniques and Digital Signatures

Akshay Atam
Stevens Institute of Technology
aatam@stevens.edu

May 27, 2024

Abstract

Open foundation models such as GPT-3 and Llama 2 have transformed artificial intelligence by making advanced AI capabilities widely accessible. These models, characterized by their large-scale and pre-trained nature, enable a variety of applications from chatbots to creative content generation. However, their openness presents significant challenges, including susceptibility to misuse for disinformation, lack of robustness, and ethical concerns such as bias and privacy violations. This review paper explores various Natural Language Processing (NLP) methodologies that can address these limitations. Additionally, it proposes a novel policy of embedding "model signatures" within the outputs of these models to enhance traceability and accountability. By leveraging explainability techniques, automated fact-checking, and ethics-aware fine-tuning, this study aims to fortify open foundation models against misuse and enhance their ethical deployment.

Keywords Open Foundation Models · Natural Language Processing (NLP) · Explainability · Automated Fact-Checking · Ethics-Aware Fine-Tuning · Model Signatures · Traceability · Accountability · AI Governance · Disinformation · Bias Mitigation · Robustness · Ethical AI · Cryptographic Hash Functions · Invisible Watermarking

1 Introduction

Open foundation models, such as GPT-3 [1] and Llama 2 [2], represent a paradigm shift in the field of artificial intelligence. These models, characterized by their large-scale and pre-trained nature, have democratized access to advanced AI capabilities, enabling a wide range of applications from chatbots to creative content generation. Their open nature, with publicly available model weights, fosters innovation by allowing researchers and developers to build upon existing models, promoting a collaborative ecosystem.

However, this openness also brings significant challenges. The widespread availability of these powerful models raises concerns about their misuse for generating disinformation, cyberattacks, and other malicious activities [3]. Additionally, there are technical limitations, such as the lack of robustness and interpretability, as well as ethical and operational concerns, including bias and the difficulty of ensuring compliance with regulations. These challenges necessitate a balanced approach that maximizes the benefits of open foundation models while mitigating their risks.

The primary aim of this paper is to explore various Natural Language Processing (NLP) methodologies that can address the limitations of open foundation models. In addition, it proposes a novel policy of embedding "model signatures" in the outputs of these models. These signatures are intended to provide traceable information about the model and its provenance, facilitating better management of disinformation and illegal content.

The rest of the paper are as follows: Section 2 and 3 provides a background and limitations on open foundation models, respectively. Section 4 provides methodologies, possible solutions, to address the limitations of open foundation models. Section 5 address the model signature policy recommendation. The report concludes with discussion in Section 6, followed by acknowledgement in Section 7.

2 Background on Open Foundation Models

A foundation model is a model trained at broad scale that can be adapted to a wide range of downstream tasks. Notable examples of foundation models include BERT [4], GPT-3 [1], and CLIP [5]. While the underlying technology of foundation models—self-supervised learning with neural networks [6]—has been around for decades, the scale and capabilities of these recent models represent a significant advancement.

The novelty of foundation models lies in the unprecedented scale of these models and their ability to perform tasks beyond their initial training regime. Two key concepts underscore the significance of foundation models: emergence and homogenization.

Emergence refers to the phenomenon where complex patterns, behaviors, or properties arise from the interactions of simpler components within a system. In other words, emergence is the process through which novel and often unexpected characteristics or structures emerge at a higher level of organization that cannot be predicted solely by understanding the individual components. This concept is commonly observed in various fields such as physics, biology, sociology, and artificial intelligence, where the interactions of individual elements give rise to collective behaviors or properties that are not present in the individual parts [7].

Homogenization, on the other hand, refers to the process of making something uniform or standardized. In the context of artificial intelligence and machine learning, homogenization can occur when diverse inputs or perspectives are reduced to a more uniform output or representation. This can lead to a lack of diversity, bias, or exclusion in the outcomes generated by AI systems. Homogenization can have negative implications, such as reinforcing existing biases, limiting creativity, or overlooking important variations in data that could lead to more inclusive and accurate results [7].

These concepts are part of the broader narrative of AI development [8]. Machine learning (ML) itself involves learning how to perform tasks from examples, with deep learning (DL) representing a subset where the system learns the features needed for prediction. Foundation models extend this idea by enabling the emergence of new functionalities that go beyond feature extraction.

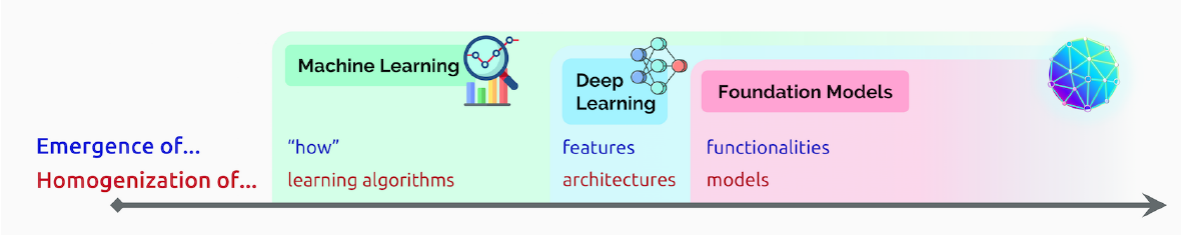

Figure 1: With machine learning, how a task is performed emerges from examples with deep learning, which fits within machine learning, what emerges are the set of features used to perform prediction, while with foundation models, what emerges are new functionalities like in context learning. With regard to the second concept, machine learning, represents a step towards homogenization of learning algorithms, since a single algorithm, such as logistic regression, can power a wide range of applications in deep learning. Rather than designing separate features for each application, the same architecture can be used with foundation models. [8]

The concept of homogenization is exemplified by the evolution from traditional ML algorithms, like logistic regression, which can be applied to various tasks, to DL architectures, where the same model can serve multiple applications [7]. Foundation models further this trend by providing a unified model capable of handling a diverse array of tasks without the need for separate designs.

In the context of foundation models and AI systems, emergence can refer to the unexpected behaviors or capabilities that arise from the interactions of the model's parameters and training data. Homogenization, on the other hand, can refer to the potential risk of standardizing outputs or perspectives in a way that limits diversity and inclusivity in the results generated by these models [8]. Understanding both emergence and homogenization is crucial for developing AI systems that are ethical, robust, and capable of producing diverse and equitable outcomes.

2.1 Current Policies and Regulations

The regulatory landscape for foundation models is evolving [8], with various approaches being considered across different regions:

1. **United States:** The U.S. Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence emphasizes the need for transparency and accountability in AI. However, there is ongoing debate about how to balance innovation with regulation.
2. **European Union:** The EU's AI Act proposes stringent regulations for high-risk AI applications, which could impact the development and deployment of foundation models. Models with fewer than 10^{25} floating point operations may be exempted, highlighting the need for nuanced regulation.
3. **United Kingdom:** The UK's AI Safety Institute focuses on both open and closed models, aiming to develop best practices for AI safety and security.
4. **Industry Guidelines:** Organizations like the Partnership on AI have introduced guidelines for the responsible deployment of foundation models, advocating for transparency and the avoidance of open release for highly capable models.

3 Limitations of Open Foundation Models

Open foundation models, despite their numerous benefits, face several significant limitations that can hinder their effectiveness and pose risks to society. These limitations can be broadly categorized into technical, ethical, and operational challenges.

3.1 Technical Limitations

Lack of Robustness. Open foundation models often lack robustness, making them vulnerable to adversarial attacks and perturbations. Adversarial examples—small, intentional modifications to input data—can cause these models to produce incorrect or harmful outputs. This vulnerability can be exploited to generate misleading information or manipulate model outputs in malicious ways.

Interpretability and Explainability. One of the primary technical challenges is the lack of interpretability and explainability. These models function as black boxes, making it difficult to understand how they arrive at specific decisions or outputs. This opacity hinders trust and accountability, as users cannot easily discern whether the model's outputs are based on sound reasoning or are biased and flawed.

Bias and Fairness. Open foundation models are trained on large datasets that often contain inherent biases. As a result, the models can learn and perpetuate these biases, leading to unfair or discriminatory outcomes. This is particularly concerning in applications such as hiring, lending, and law enforcement, where biased decisions can have significant societal impacts.

3.2 Ethical and Social Concerns

Misuse and Malicious Use The openness of these models makes them susceptible to misuse. Malicious actors can use them to generate disinformation, deepfakes, and other harmful content. For example, sophisticated language models can produce persuasive fake news articles or social media posts that can spread misinformation and influence public opinion. Similarly, text-to-image models can create realistic but fake images or videos, leading to the spread of non-consensual intimate imagery or child sexual abuse material (CSAM) [3].

Privacy Violations Training data for open foundation models often includes large amounts of publicly available text, which can inadvertently include personal and sensitive information. Models can sometimes memorize and inadvertently reproduce this information, leading to privacy violations. This is especially problematic when the model is used to generate outputs that contain sensitive data about individuals without their consent.

Ethical Decision-Making These models lack a nuanced understanding of ethical considerations. They may produce outputs that are ethically questionable or outright harmful. For instance, a model generating content for customer service might inadvertently provide advice that is dangerous or unethical due to its lack of contextual understanding.

3.3 Operational Challenges

Scalability and Maintenance Maintaining and scaling open foundation models requires significant computational resources and infrastructure. As the models grow larger and more complex, the cost of training and deploying them increases. This can be a barrier for smaller organizations that lack the necessary resources to maintain these models effectively.

Compliance and Regulation Ensuring compliance with varying global regulations is another operational challenge. Different countries have different standards and laws regarding data privacy, content moderation, and AI ethics. Developers of open foundation models must navigate this complex regulatory landscape to ensure their models comply with all relevant laws and guidelines. This is particularly challenging for models used across multiple jurisdictions.

Version Control and Provenance Tracking the lineage and versions of open foundation models is critical for accountability and improvement. However, given their open nature, it is challenging to maintain proper version control and provenance. This makes it difficult to identify which version of a model was used to generate a specific output, complicating efforts to trace back and address issues that arise from the use of these models.

4 Methodologies to Address Limitations

We highlight three potential domains where existing methodologies could be used to address the limitations and challenges faced by open foundation models. The issues pertaining to open foundation models have been adapted from the policy brief of Bommasani et al. [3]. These include eXplainability AI (XAI) techniques, fact-checking, and ethics aware fine-tuning.

4.1 eXplainability AI (XAI) Techniques

Existing research has shown that the performance of language models is strongly dependent on the scale and less on model's shape [9]. Model explainability is crucial to address the opacity of modern deep neural networks [10, 4, 11]. Lundberg et al. [12] proposes SHapley Additive exPlanations (SHAP), a unified local-interpretability framework with a rigorous theoretical foundation on the game-theoretic concept of Shapley values [13]. SHAP values are based on the idea that the output produced by each possible combination of features (where each feature can be used or not be used by the model) should be considered to determine their importance. This means that a separate model should be trained for each possible combination of the available features, always with the same hyperparameters and training data.

In addition to using Shapley values for model explainability, LIME (Local Interpretable Model-agnostic Explanations) [14] is also a widely used technique for post-hoc model explainability. LIME provides local interpretability for specific instances by approximating complex model's prediction with simpler, interpretable models [15]. The weights serve as feature important scores which are more robust against adversarial perturbations than gradient-based methods.

Current literature supports the use of SHAP and LIME on NLP interpretability [16] and their uses on Large Language Models (LLMs) [11]. Heyen et al. [17] evaluate the explanations using faithfulness [18] and plausibility [18]. Their work indicate that larger models provide more faithful explanations but does not align with human-generated explanations, particularly in capturing higher-level reasoning and token dependencies. Thus, there is a need for more coherent and expressive explainability metrics for LLMs; one which is objective, scalable, and human-interpretable. Alternatively, the research could be expanded on different NLP tasks, including those optimized with Reinforcement Learning from Human Feedback (RLHF).

4.2 Fact-checking

The rise of information and the lack of strict policies to combat the spread of information has lead to misinformation and disinformation on social media. Moreover, open foundation models have a higher risk of generating persuasive disinformation [3], leading to conflicts and manipulation in groups of people. Manually verifying fact checkers like PolityFact and FactCheck.org provide a solution of verifying claims based on different sources of evidence, but they are insufficient with the speed of updated information on social media. In contrast, automatic fact-checking methods rely on four different sub-tasks [19]

- **Claim Detection:** Involves identifying statements or rumors that are worth fact-checking.
- **Evidence Retrieval:** Aims to find information beyond the claim to indicate its veracity.
- **Verdict Prediction:** Determines the truthfulness of the claim based on the evidence.
- **Justification Production:** Involves explaining the reasoning behind the verdict, often using attention weights, decision-making processes, or textual explanations.

Numerous studies have focused on improving the efficacy of these systems through diverse methodologies. One notable approach involves the application of fast dot product indexing in evidence retrieval, which significantly enhances the speed and accuracy of matching claims with relevant evidence [20, 21]. This indexing technique allows for rapid and efficient retrieval of pertinent information, thereby improving the overall efficiency of the fact-checking process.

Another critical area of advancement is the development of new reasoning methods tailored for verification tasks. These novel reasoning frameworks provide more sophisticated analytical capabilities, enabling the system to better handle the complexities involved in verifying claims [22]. By incorporating advanced reasoning techniques, fact-checking systems can offer more reliable and nuanced assessments of the veracity of information.

Among the four sub-tasks mentioned above, evidence retrieval and verdict prediction are especially critical. They form the foundation for determining ground-truth information and the inference mechanisms necessary for verifying information. Existing research comprises on traditional, non-neural approaches [23, 24, 25] and neural-based approaches [26, 27]. Additionally, BERT-based language models were used on evidence retrieval into open-domain question-answers [26] and fact-checking tasks [28] and cross-encoder architecture was used for ranking multi-stage documents [29, 30] and passage re-ranking [31]. In addition, kernel-based [32] and graph-based [33] approaches provides conditions for verifying information on more complex claims [34, 35].

In recent years, numerous pre-trained language models have been developed to enhance fact-checking capabilities. Notable examples include RoBERTa [36], ALBERT [37], and XLM-R [38], which are multilingual pre-trained language models. These models offer significant potential for improving fact-checking processes. However, further research is needed across all four domains of automatic fact-checking—claim detection, evidence retrieval, verdict prediction, and justification production. Additionally, there is a need for the creation of datasets that contain up-to-date information to ensure the continued effectiveness and relevance of fact-checking systems.

4.3 Ethics Aware Fine-Tuning

Embedding ethical guidelines directly into the training data of AI models, a process known as ethics-aware fine-tuning, is a proactive approach to mitigate biases and promote fairness in AI outputs. This technique involves integrating ethical considerations and guidelines into the data and training processes, ensuring that models learn to adhere to these principles from the outset. The approach looks like the following:

1. **Data Annotation:** Data used for training is annotated with ethical guidelines in mind. This includes labeling data that may involve sensitive content or potential biases and providing clear guidelines for ethical usage.
2. **Bias Mitigation:** During the training phase, techniques such as re-sampling, re-weighting, and data augmentation are applied to reduce biases. This ensures that the model does not learn and perpetuate discriminatory patterns.
3. **Ethics-Aware Fine-Tuning:** Models are fine-tuned on datasets that have been carefully curated to reflect ethical considerations. This process includes iterative adjustments and validation to ensure adherence to ethical guidelines.

Embedding ethics in AI systems offers several benefits and challenges. Immediate response capabilities enable swift action to address misuse, preventing the spread of harmful content and mitigating potential damage. Continuous improvement is facilitated through feedback loops, ensuring that the monitoring system evolves and adapts to new threats, enhancing its effectiveness over time. Additionally, real-time monitoring provides transparency in the use of AI models, fostering accountability and trust among users. However, implementing these systems at scale poses significant challenges, particularly for large organizations with extensive AI deployments. Balancing sensitivity and specificity to minimize false positives while ensuring that true instances of misuse are detected is another critical challenge. Furthermore, ensuring that monitoring systems respect user privacy and comply with data protection regulations adds an additional layer of complexity.

While embedding ethics in AI models might seem far-fetched, this is not an unprecedented endeavor. This approach draws inspiration from Isaac Asimov's "Three Laws of Robotics," an ethical system developed for interactions between humans and robots. First introduced in Asimov's 1942 short story "Runaround," these laws have significantly influenced both science fiction and discussions on AI ethics and safety. Asimov's Three Laws state that,

- A robot may not injure a human being.
- A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Although Asimov's laws are a fictional framework and not a comprehensive set of rules for AI systems, they provide a foundation for considering the ethical implications of AI development. By embedding ethical guidelines directly into AI models, we can begin to address the complexities of real-world applications, ensuring the safety, well-being, and

autonomy of human users. Researchers, developers, and policymakers must collaborate to establish comprehensive guidelines and regulations that reflect these ethical considerations.

5 Policy Recommendation - Model Signature

This section outlines the proposal of using a "model signature" to address the critical issues related to accountability, traceability, and the prevention of misuse of open foundation models. Model signatures are unique, invisible identifiers embedded within the outputs of AI models. These signatures serve as digital fingerprints that provide traceable information about the model, including its version, the owner, and other relevant metadata. The primary objectives of this policy are to enhance accountability, facilitate the detection of misuse, and ensure the responsible deployment of open foundation models.

5.1 Key Elements

The policy of model signature must adhere to the SUIT property, providing Security, Uniqueness, Invisibility, and Traceability.

- **Security:** Model signatures should be tamper-proof and resilient to removal attempts.
- **Uniqueness:** Each model signature is unique to a specific model and version, to allow for precise identification.
- **Invisibility:** The signatures are embedded in a way that does not alter the perceived quality or meaning of the output.
- **Traceability:** The signatures enable tracing back to the model and dataset used, providing a clear provenance for generated content.

5.2 Related Work

Model cards, introduced by Mitchell et al. (2019), serve as a framework for the transparent documentation of machine learning models [39]. They provide detailed evaluations across diverse demographic groups and conditions, enhancing accountability and traceability in AI systems. The concept of model cards has evolved to include additional aspects such as consumer labels for ML models, principles for explainable models, and the incorporation of environmental and financial impact considerations [40].

The importance of model cards in promoting ethical AI practices is evident from their widespread adoption and subsequent adaptations. For instance, the integration of fairness analysis in model cards aims to highlight potential biases and mitigate their impact, particularly in sensitive domains like healthcare [41]. Expanding the scope of bias reporting to include non-social factors such as disease-dependent and anatomic factors has been proposed to ensure comprehensive bias analysis and safer deployment of AI models in clinical settings.

While model cards provide a structured way to document model characteristics and performance, the proposed "model signature" policy aims to embed metadata directly into model outputs. This metadata, or signature, could include information about the model, its owner, and the dataset used. Such a signature would enhance the traceability and accountability of model outputs, making it easier to identify the source of disinformation or illegal content.

Combining model signatures with model cards could provide a robust framework for AI governance. While model cards offer detailed documentation accessible to developers and regulators, model signatures ensure that every output can be traced back to its source, providing an additional layer of security and accountability. This dual approach can help address both the transparency and traceability challenges in deploying open foundation models [39].

5.3 Technical Implementation

5.3.1 Invisible Watermarking

Invisible watermarking techniques have been developed to embed identifiable markers within images that are imperceptible to the human eye but detectable through specific algorithms. SynID, used in models like Imagen [42], effectively incorporates these invisible watermarks, making synthetic images identifiable while maintaining their visual integrity. Additionally, researchers are exploring the possibility of embedding watermarks directly into generative models, further enhancing the detectability of synthetic images [43].

In the domain of NLP, invisible watermarking would involve embedding subtle, imperceptible variations in the text that can be detected by specialized algorithms. These variations can include slight changes in synonym usage, punctuation, or sentence structure that are not noticeable to human readers but can be identified computationally.

5.3.2 Cryptographic Hash Functions

A hash function is an algorithm that transforms an input value into a fixed-length output value, often referred to as a "hash" [44]. Cryptographic hash functions can generate a unique hash value based on the model and dataset. This hash value can be embedded in the output as a signature. The work by Cohen et al. [44] provide the following key takeaways:

1. Non-linear cryptographic hash functions, such as SHA, possess error correction capabilities comparable to systematic random linear codes.
2. The authors combine error correction and hash verification, offering a practical solution for reliable data transmission over noisy channels.
3. The theoretical and empirical results demonstrate that non-linear cryptographic hash functions can achieve capacity in the asymptotic regime, ensuring low probability of error in data transmission.
4. The findings open new avenues for the use of non-linear cryptographic hash functions in applications requiring both data authenticity and error correction.

By embedding model signatures using non-linear cryptographic hash functions with robust error correction capabilities, this AI policy framework enhances the security, reliability, and integrity of AI-generated outputs. It provides a practical approach to authenticating AI models, ensuring their outputs remain trustworthy and resistant to tampering in real-world applications.

5.4 Case Study

Note: Although the invisible watermarking technique, non-linear cryptographic hash functions, and ethics-aware fine-tuning have not been implemented in practice and are presented here hypothetically, this case study is designed to showcase the potential importance and benefits of a model signature policy. The following scenario particularly addresses the risk of disinformation in open foundation models from the policy brief of Bommasani et al. [3] and illustrates how these techniques could enhance accountability and traceability in AI-generated content.

Background

Horizon is a new social media platform where users share their innovative AI-generated text and images, based on real scenarios. It consists of diverse forums where users are able to post their thoughts on news of the world. Currently, Horizon is facing challenges with disinformation and other harmful content through their posts and comments. To address this, the AI team at Horizon decides to implement the model signature policy using invisible watermarking technique and non-linear cryptographic hash functions.

Objectives

- To embed unique, invisible signatures in AI-generated content to trace its origin.
- To make sure that the signatures do not affect the user experience or the readability of the content.
- Enable robust detection and verification of signatures even after content modifications.

Implementation Steps

1. **Invisible Watermarking**
    - **Selection of Watermarking Points:** The platform identifies suitable locations within the text (or images) where subtle changes can be made without altering the content's meaning. These include slight variations in punctuation, synonym choices, or sentence structure are considered.
    - **Embedding the Watermark:** Using an advanced watermarking algorithm, unique identifiers are embedded in these selected points. These identifiers are imperceptible to human readers but detectable by specialized software.
    - **Verification Tool:** Develop a tool capable of scanning AI-generated content to detect the embedded watermarks. This tool uses a reference database to match the detected watermark with the original AI model and its version.

2. **Non-Linear Cryptographic Hash Functions**
    - **Hash Generation:** Every time an output is generated by the LLM, it generates a unique cryptographic hash based on its parameters, training data, and version. This hash is created using a non-linear cryptographic function, making it resilient to tampering.
    - **Embedding the Hash:** The cryptographic hash is embedded within the content using steganographic techniques. This process ensures that the hash is hidden within the text and remains robust against text modifications.
    - **Detection and Verification:** A separate verification tool is developed to extract and verify the cryptographic hash from the content. This tool cross-references the extracted hash with the platform's database to confirm the model's identity and provenance.

Implementation

- Horizon's AI team trains a new large language model, *Dodo*, using a diverse dataset. Ethical guidelines are embedded into the training data through ethics-aware fine-tuning. Upon completion, a unique non-linear cryptographic hash is generated for Dodo.
- The AI team integrates the invisible watermarking algorithm and hash embedding process into Dodo. Each time Dodo generates a post (or comment), it embeds an invisible watermark and the cryptographic hash into the text. These signatures do not affect the content's readability or user experience.
- As Dodo generates content, Horizon's real-time monitoring system continuously scans for embedded watermarks and cryptographic hashes. When a suspicious post is detected, the system extracts the watermark and hash, verifying the content's origin and ensuring it complies with ethical guidelines. If the post (or comment) does not comply with the standards, the specific post (or comment) gets deleted and a flag is imposed on the user.

Incident

A post containing disinformation about a public health issue linking to a minority group is flagged by users. Dodo's monitoring system scans the post, detects the embedded watermark and cryptographic hash, and traces it back to the user.

Response

The flagged post is removed swiftly, and Horizon's moderation team investigates the incident. The traceability provided by the watermark and hash allows the team to review the model's output and take corrective actions on the user. This includes updating the ethical guidelines and refining the training data to prevent similar incidents in the future.

Outcome

The implementation of the model signature policy significantly enhances the platform's ability to manage and moderate AI-generated content. The invisible watermarking and non-linear cryptographic hash functions ensure robust traceability and accountability, fostering a safer and more trustworthy online environment for Horizon's users.

6 Discussion

The findings from this study contribute to the ongoing discourse on managing the risks and benefits associated with open foundation models in AI. This study contributes to the body of knowledge by presenting a comprehensive review of the limitations inherent in open foundation models and proposing robust methodologies to address these challenges.

One significant contribution is the exploration of explainability techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). These techniques are critical in demystifying the decision-making processes of large language models, thereby enhancing transparency and trust. The review highlights that while current explainability methods provide valuable insights, there is a need for more advanced metrics that can capture higher-level reasoning and token dependencies specific to large language models.

Another key contribution is the focus on automated fact-checking to combat the spread of misinformation. The paper underscores the importance of claim detection, evidence retrieval, verdict prediction, and justification production in developing robust fact-checking systems. By integrating fast dot product indexing and advanced reasoning methods, the study suggests improvements that could significantly enhance the efficacy and reliability of automated fact-checking systems.

The proposal of embedding ethics directly into AI models through ethics-aware fine-tuning represents a proactive approach to mitigating biases and promoting fairness. This approach, inspired by the principle of Asimov's Three Laws of Robotics, seeks to ensure that models adhere to ethical guidelines from the outset. The discussion emphasizes that while embedding ethics into AI models is complex, it is essential for creating systems that are not only technically proficient but also socially responsible.

The introduction of a "model signature" policy provides a novel approach to enhancing accountability and traceability of AI-generated content. The discussion details how invisible watermarking and cryptographic hash functions can be utilized to embed unique, tamper-proof identifiers in AI outputs. This policy aims to prevent misuse by enabling precise identification of the model and version used to generate specific outputs, tracing it back to its origin, thus facilitating accountability and compliance with ethical guidelines.

In summary, this study provides a multifaceted approach to addressing the challenges posed by open foundation models. By leveraging advanced NLP methodologies, it proposes a balanced framework that enhances the benefits of these models while mitigating their risks. The implementation of explainability techniques, automated fact-checking, ethics-aware fine-tuning, and model signatures collectively contributes to the development of AI systems that are ethical, accountable, and transparent. Future research should focus on understanding these methodologies and exploring their practical applications to ensure the responsible deployment of open foundation models.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[3] Kevin Klyman Shayne Longpre Ashwin Ramaswami Daniel Zhang Marietje Schaake Daniel E. Ho Arvind Narayanan Percy Liang Rishi Bommasani, Sayash Kapoor. Considerations for Governing Open Foundation Models. `https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models`, 2023.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[6] Virginia De Sa. Learning classification with unlabeled data. *Advances in neural information processing systems*, 6, 1993.

[7] Samuel Albanie. On the Opportunities and Risks of Foundation Models (intro). `https://www.youtube.com/watch?v=ZshcPdavsdU&t=218s`, 2023.

[8] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun

Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021.

[9] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[10] Edoardo Mosca, Maximilian Wich, and Georg Groh. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, 2021.

[11] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. arxiv. *arXiv preprint arXiv:2108.04840*, 2021.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[13] Lloyd S Shapley et al. A value for n-person games. 1953.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[15] Anshul Goel. Model Explainability using SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). `https://medium.com/@anshulgoel991/model-exploitability-using-shap-shapley-additive-explanations-and-lime-local-interpretable-cb4f559` 2023.

[16] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pages 4593–4603, 2022.

[17] Henning Heyen, Amy Widdicombe, Noah Yamamoto Siegel, Philip Colin Treleaven, and Maria Perez-Ortiz. The effect of model size on llm post-hoc explainability via lime. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024.

[18] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

[19] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[20] Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, 2023.

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[22] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*, 2023.

[23] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[24] Xuming Hu, Zhijiang Guo, Guanyu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. Chef: A pilot chinese dataset for evidence-based fact-checking. *arXiv preprint arXiv:2206.11863*, 2022.

[25] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.

[26] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

[27] Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932*, 2020.

[28] Chris Samarinas, Wynne Hsu, and Mong Li Lee. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91, 2021.

[29] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*, 2019.

[30] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*, 2019.

[31] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[32] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Fine-grained fact verification with kernel graph attention network. *arXiv preprint arXiv:1910.09796*, 2019.

[33] Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Reasoning over semantic-level graph for fact checking. *arXiv preprint arXiv:1909.03745*, 2019.

[34] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics, 2019.

[35] Michael Schlichtkrull, Vladimir Karpukhin, Barlas Oğuz, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. Joint verification and reranking for open fact checking over tables. *arXiv preprint arXiv:2012.15115*, 2020.

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[37] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[38] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[39] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[40] DeBrae Kennedy-Mayo and Jake Gord. " model cards for model reporting" in 2024: Reclassifying category of ethical considerations in terms of trustworthiness and risk management. *arXiv preprint arXiv:2403.15394*, 2024.

[41] Carolina AM Heming, Mohamed Abdalla, Monish Ahluwalia, Linglin Zhang, Hari Trivedi, MinJae Woo, Benjamin Fine, Judy Wawira Gichoya, Leo Anthony Celi, and Laleh Seyyed-Kalantari. Benchmarking bias: Expanding clinical ai model card to incorporate bias reporting of social and non-social factors. *arXiv preprint arXiv:2311.12560*, 2023.

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[43] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

[44] Alejandro Cohen and Rafael GL D'Oliveira. Error correction capabilities of non-linear cryptographic hash functions. *arXiv preprint arXiv:2405.01495*, 2024.