Beyond Technical Transparency: Explainability as a Safeguard Against Manipulative AI*

Ermanno Petrocchi
University of Macerata
e.petrocchi1@unimc.it

Simona Tiribelli
University of Macerata
ITGH,
PathCheck Foundation
simona.tiribelli@unimc.it

Berk Buzcu
HES-SO Valais-Wallis
berk.buzcu@hevs.ch

Elia Pacioni
HES-SO Valais-Wallis
University of Extremadura
elia.pacioni@hevs.ch

Davide Calvaresi
HES-SO Valais-Wallis
davide.calvaresi@hevs.ch

***Abstract*—Large Language Models (LLMs) now write with a fluency and persuasiveness that can subtly steer users' choices. When their outputs lack clear and comprehensible explanations, this persuasive power risks undermining human decision-making capacity, raising serious ethical concerns. Current explainable artificial intelligence (XAI) techniques focus primarily on technical transparency for epistemic purposes (how a model works); they are rarely intended to reveal to the user the kind of influence they are subject to. Drawing on the Indifference View of manipulation, we advance a preliminary framework that reconceives explainability as both an epistemic and an ethical imperative. The core idea is based on explanatory metadata: layered annotations that accompany model outputs with four complementary types of explanation-informative, justificatory, causal, and precautionary-which give models the ability to detail the reasons underlying the influence they exert. Doing so shifts the XAI goal from mere transparency to responsible influence. It positions explanations as a safeguard against the manipulative behavior of generative AI systems, laying the groundwork for future methods that measure, audit, and actively constrain ethically problematic influence.**

***Index Terms*—Explainable AI, Manipulation, AI Ethics**

I. Introduction

In recent years, large language models (LLMs) have reached a level of sophistication that makes them extremely persuasive tools. They do not merely provide coherent responses to textual inputs, but are also capable of inferring and predicting user preferences based on linguistic patterns and past interactions, detecting emotional signals, and dynamically adapting responses to satisfy explicit or implicit requests [1], [2]. Consequently, interaction with LLMs is neither neutral nor purely informative: it can take the form of profound influence that often goes beyond the user's awareness.

What makes this influence particularly problematic is that it is exercised without being made explicit. Specifically, LLMs operate without necessarily clarifying why their influence occurs, i.e., without revealing the rationale behind their responses. Indeed, even when LLMs provide explanations for their outputs, these typically remain semantic constructions rather than faithful and reliable reflections of the internal processes and logic used to generate the output. This risks making their influence manipulative. It is therefore crucial to ask whether it is legitimate for such models to guide users' thoughts and behaviors without any substantial and trustworthy attempt to make explicit their means of influence.

In this context, explainability plays a central role. Traditionally conceived as a response to the need to understand how a system works, today, in light of the significant influence that LLMs can exert, it must also be understood as a tool to ensure that models' influence remains beneficial. Explainability is no longer simply a means to verify model outputs but a necessary condition to make the rationale behind AI-driven influence explicit, supporting beneficial and respectful interactions.

Explainable AI (XAI) techniques can serve to meet this need. These techniques aim to make the decision-making processes of AI systems more comprehensible, translating, at least in part, the complexity of their internal mechanisms into representations or explanations that are accessible to humans. XAI aims to improve users' confidence in the system and enable a more critical evaluation of the answers provided. However, XAI techniques suffer from two main limitations: a technical one, related to their ineffectiveness in making LLMs' internal mechanisms truly intelligible, offering only partial explanations, and an ethical one, due to the lack of explanations aimed at providing reasons for the influence exerted on users. When exerted without disclosing its underlying reasons, LLMs' influence fails to meet the normative principle that should guide human interactions and risks turning into harmful manipulation. Therefore, this asymmetry between technical explanations and ethical justifications leaves uncovered a fundamental knot in human-LLMs interaction. When a system can shape beliefs and behaviors without providing meaningful reasons for its influence, its interaction risks becoming manipulative, regardless of the system's intentions or the user's awareness.

Therefore, overcoming the limitations of existing XAI techniques becomes essential. What is needed is not just more transparency about models' functioning but a form of explainability capable of making the influence of LLMs ethically acceptable. Starting from the view of manipulation as indifference to reasons – a leading normative account in AI ethics – this paper proposes a preliminary framework

This work has been funded by the European Union - NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem grant ECS00000041 - VITALITY - CUP D83C22000710005, and the Spanish Ministry of Economy and Competitiveness grant PID2023-147409NB-C22, and the Junta de Extremadura (GR24142).

for alternative explainability methods, i.e., based on reasons. These have the primary objective of ensuring that the influence of LLMs is reason-revealing, thus preventing harm. In doing so, the goal is to align the model behavior with the standards for respectful and beneficial interactions.

The remainder of the paper is structured as follows. Section II introduces, the theory that allows the evaluation of LLMs' influence, namely the Indifference View. Section III elaborates on LLM as black boxes. Section IV exposes the limitations of current XAI techniques. Section V outlines preliminarily new possible explainability methods, placing them within the existing landscape of XAI techniques. Section VI offers a technical roadmap for implementing the proposal. Section VII concludes.

II. Explainability and manipulation: The Indifference View

Explainability aims to make the process through which an AI system, such as an LLM, computes a given input and generates a specific output more transparent. To date, explainability serves both technical and epistemic functions. Technically, it aims to clarify the inner workings of a model, e.g., through the attribution of features or the creation of attention maps. Epistemically, it fosters human understanding by enabling the verification of whether a system functions correctly and whether flaws, biases, or inconsistencies are present. Based on that, humans can decide whether and to what extent to trust or follow algorithmic suggestions. In this sense, explainability promotes more informed and responsible engagement with AI.

However, this epistemic value alone does not fully capture the normative significance of explainability. As discussed in the Introduction, the capacity of LLMs to subtly influence users raises ethical concerns that extend beyond mere understanding. This section argues that explainability should not be limited to enhancing user comprehension; it should express a broader ethical commitment to responsible forms of influence. This marks a crucial shift: while understanding refers to the user's capacity to know how the system works, ethical justification demands that the system disclose the reasons underlying its actions, especially when it exercises an influence. Therefore, it is not enough for a model to be understandable; it must also provide reason-based explanations that make its influence legitimate and beneficial.

To address this concern, this work draws on the theory of manipulation as outlined by the *Indifference View* [3], [4]. This theoretical lens has been chosen as it proves particularly useful in assessing LLMs' explainability, as it identifies manipulation not just in terms of its effects or methods alone, but also through the lens of absence of concern for reason-giving in the design choice of influence strategies.

Before delving deeper into this view, it is important to acknowledge that manipulation is a dynamic and multifaceted concept that resists definitive characterization. The various existing accounts highlight certain manipulation features while leaving others in the background. Some focus on outputs (e.g., violation of autonomy) or manipulative methods (covertness or bypassing rationality) and although they stress important aspects, fall victim to counterexamples and inconsistency that question their validity [5]. The Indifference View is not exhaustive as well. Yet, by focusing on the norm that should guide ideal interactions, namely the revealing-reason concern, it helps clarify how LLMs should behave to meet that normative standard.

Following the Indifference View, what marks a manipulative influence is: (i) being goal-directed, thereby excluding accidental or incidental forms of influence; (ii) the chosen means of influence are not selected to reveal reasons to the target. From this perspective, the crucial aspect is the manipulator's indifference towards providing the manipulated person with the reasons behind the influence they exert. Manipulation is fundamentally about a failure to respect the normative expectations embedded in interpersonal (or human–AI) relations, namely the duty to offer reasons for one's influence.

Applied to AI systems, it is essential to highlight that even when an LLM provides correct or helpful answers, its influence can be manipulative if reason-revealing explanations do not accompany it. In other words, a formally valid output is not enough to rule out manipulation: what matters is whether the system's influence reflects a concern for making its rationale apparent. Therefore, LLMs' influence is problematic not just because obscure mechanisms may unknowingly guide users, but because their influence can be exercised without concern for providing reasons. In such cases, LLMs bypass the requirement to reveal their influence, violating in this way the standards of respectful, i.e., reason-sensitive, interaction.

From this perspective, explainability becomes a fundamental ethical safeguard. When the reasons behind AI systems' output are withheld, the reason-revealing requirement is not satisfied, and then AI-generated influence risks becoming illegitimate and potentially harmful.

III. LLM as Black Boxes

LLMs are widely used in modern Natural Language Processing (NLP) applications due to their high performance in text generation, synthesis, and translation [6]. However, their opacity presents a significant ethical challenge, prompting researchers to investigate the underlying mechanisms that contribute to the black-box nature of LLMs. The causes of this black-box nature can be divided into three main categories: architecture, training process, and post-training operations. From an architectural point of view, LLMs are based on Transformers, complex models with hundreds of billions of parameters [6], [7] that process information through attention mechanisms and distributed representations that are difficult to interpret [8]. Attention matrices and semantic abstractions learned in the upper layers do not always allow us to understand how a specific output is produced [9]. Training involves using large amounts of diverse and often unstructured data [10], and a lack of transparency in data selection [11] introduces biases that are difficult to detect and correct. Post-training operations, such as quantization [12], pruning [13],

and fine-tuning [14], modify the model's behavior to improve its efficiency or adapt it to specific contexts, but often without being able to precisely track the impact of such modifications on the outcomes produced. Furthermore, methods such as Reinforcement Learning from Human Feedback (RLHF) [15], while improving the quality of responses, add layers of interpretation complexity related to the optimization algorithms deployed. Finally, due to their probabilistic operation, the nondeterministic nature of LLMs implies that the same input can generate different outputs [16], making it difficult to validate and replicate the outputs, especially in sensitive areas.

Overall, LLMs are powerful but inherently opaque due to structural complexity, limited transparency of training data, and variable optimization. Although several XAI techniques have been developed to address these issues, their effectiveness remains limited.

IV. Limitations of XAI techniques

XAI techniques developed so far represent a vital attempt to improve the transparency of AI systems, and LLMs in particular. However, these techniques are insufficient to guarantee truly effective and meaningful explainability for users, for technical and ethical reasons.

From a technical standpoint, XAI techniques are often designed to make AI models interpretable. However, they rarely successfully explain the mechanisms governing the outputs of LLMs in a complete and comprehensible manner. Popular methods, such as attention visualizations, saliency maps, or feature importance scores, often provide only partial, complex, and technically dense information. Furthermore, explainability applied to GenAI, particularly to LLMs, presents unique challenges compared to traditional AI models [17]. First, many interactions with these systems are inherently interactive, dynamic, and context-dependent. In such cases, explainability cannot simply describe the model's behavior; it must also make the system's influence on the user during interaction understandable, and vice versa [18]. Second, GenAI systems rely on much larger and more complex architectures and datasets than traditional AI models. While conventional models are typically trained on closed and well-defined datasets, LLMs are built using a variety of public and private data sources and often include modular components, such as retrieval-augmented generation or code interpreters [19], [20]. As a result, the final output of a GenAI system can be the composite product of multiple interacting phases, tools, and data sources, making it challenging to provide a coherent and unambiguous explanation. Third, the outputs generated by GenAI systems are qualitatively different: whereas a traditional classifier produces a single label, generative systems produce full texts, which imply numerous choices at the semantic and stylistic levels [21].

From an ethical point of view, the critical issues with XAI techniques are even more pronounced. Many explainability tools are not designed to provide end users with relevant reasons for the influence exerted by LLMs. Rather, for internal purposes like debugging or regulatory compliance. This kind of explainability produces a technical transparency that remains completely indifferent to the type of reasons provided and that matter to foster beneficial influence. Explainability, therefore, risks being ineffective for addressing manipulation, and insufficient for promoting responsible interaction [22].

Beyond that, even more concerning, algorithmic transparency can become a direct means for manipulation. System designers, indeed, may implement explainability features that give the impression of openness and clarity, while actually failing to disclose the normative rationale behind the system's influence [4]. They could provide enough information to appear transparent while guiding users' behavior to benefit the system's operators. In this sense, so-called transparent explanations can function as manipulation tools, offering information that nudges users toward desired outcomes (e.g., continued engagement or consumption). For instance, a system may explain simply because it is supported by a library that implements a specific function. However, such an explanation would not provide any *pertinent* reason for the influence exerted by the system. In such cases, explainability risks becoming a formality that hides manipulative pitfalls.

Ultimately, current XAI techniques, while improving formal transparency, do not provide ethically adequate explanations: they do not clarify why LLMs influence users in certain ways. Without such normative reason-giving, they risk manipulative consequences even in the presence of correct or transparent outputs.

V. Alternative methods for LLMs' explainability

As argued above, despite recent efforts, XAI continues to show clear limitations. Empirical studies suggest that existing techniques offer modest benefits [23], [24], and some even define XAI research as 'largely unproductive' [25]. The advent of generative models has exacerbated the problem: LLMs generate fluid but not always factual responses, which can report misleading, inaccurate, or even harmful content (hallucinations) [26]. This requires an explainability that goes beyond technical transparency and explicitly focuses on the influence exerted on users.

To address this challenge, this paper proposes a novel, albeit preliminary, approach to explainability, centered on the idea of explanations as tools for counteracting potential manipulative influences. Rather than merely describing the inner model's functioning, this approach emphasizes the need to provide clear and meaningful reasons for the influence exerted, through various kinds of explanations. The proposal's core consists of using explanatory metadata that allows various types of explanations: justificatory, causal, informational, and precautionary [27]. Metadata-based explanations extend beyond the computational and mechanistic explanations provided by traditional XAI methods, which typically focus on isolated technical aspects, such as feature rankings or gradient-based attributions, without clarifying the underlying reasons for a system's influence. The proposed approach, by broadening the explainability range to accommodate an understanding that encompasses multiple explanatory dimensions, marks a crucial

step toward achieving influence-aware explainability.

A key example of this metadata is the logic underlying the user's input. For instance, the model could explain how it processed and interpreted the user's request, as well as the criteria it used to generate the response. This would provide the user with an informative explanation. An example might be: "This formulation is based on points X and Y in your question". This explanation would help the user understand how the model's input interpretation affected the answer. Otherwise, another kind of informative explanation could make explicit how the question's wording shaped the model answer, disclosing how different wording could have led to a different outcome. The model would provide reason-revealing explanations for its influence by making this interpretive process transparent.

In addition, explanations could be based on other metadata, such as the model's confidence level, which not only indicates the certainty of the output but also defines how much weight to assign to the information provided. For example, if a chatbot enriched its answers with statements like 'this information has a confidence level of 85%. I provided this answer because it has the highest certainty score among available options.' This statement provides a cautionary explanation, enabling the user to critically evaluate the degree of uncertainty in the output and decide whether to trust the system or seek further confirmation elsewhere.

Another type of metadata could involve paraphrasing the model's system constraints, particularly for systems involving prompt engineering. These data could provide the user with a justificatory and precautionary explanation revealing the rationale for the generated response. In an educational setting, for instance, a learning-support chatbot might be programmed to avoid directly showing homework answers. Then, if a student asked for the solution to their algebra problem, the chatbot might answer: 'Rather than giving you the solution directly, let's work through this together. I'm here to help you figure it out on your own. To solve this problem, try to isolate the variable and...' This kind of response would respect the educational constraint of not providing the solution outright without disclosing the technical configuration of the system. At the same time, it communicates the rationale behind the model's influence, making it transparent to the user why a certain suggestion or refusal is being made.

These examples show the potential of an alternative explainability, where metadata not only increase systems' transparency, but also help to counteract the risks of manipulation, clarifying how LLMs influence users. Metadata-based explanations are not mutually exclusive, but complementary dimensions of an integrated strategy oriented at making the influence exerted by LLMs reason-sensitive. In real contexts, it is indeed essential to integrate different levels of explanation (i.e., what the system does, how it acts, why it does it) to design truly reliable and normatively legitimate systems.[1]

[1]However, it should be noted that the reliability of these explanations also presents significant challenges, particularly because they are generated by the same system that exerts influence over the user. This overlap raises important ethical concerns regarding trust, deserving future further investigation.

A. Explainability through metadata within the XAI techniques landscape

This integration of metadata to provide reasons for LLMs' influence contributes to the broader debate on GenAI explainability. This proposal can be distinguished from three main common approaches:

- Interpretable models through auxiliary architectures: several methods improve LLMs' explainability by integrating external components. Some employ auxiliary models, such as graph neural networks or external knowledge bases [28], others leverage multi-model architectures or combine symbolic logic with internet-based retrieval mechanisms [29], [30]. One relevant approach is Retrieval-Augmented Generation, which binds the output to a predefined and analyzable corpus, making the responses more interpretable and verifiable [19]. This paper's proposal differs as it does not merely enhance technical traceability through external components but seeks to reveal the reasons behind the model's influence.
- Prompt-based explanations: techniques such as Chain-of-Thought prompting allow models to be guided in generating explanations step-by-step, simulating an articulated reasoning [31]. These strategies leverage the interaction between the prompt and the model's output to generate an explanation that aligns with the internal structure of the generative process. This paper's proposal stands apart by introducing layered, metadata-driven explanations that clarify the model's interpretive process, rather than simulated internal reasoning.
- Self-explaining models: this is based on LLMs' ability to produce explanations independently, without resorting to external XAI techniques. LLMs can generate personalized explanations for their outputs, explain the decisions of other models, or detect patterns in the data through self-prompting techniques [31]–[33]. This paper's proposal stands out for prioritizing ethically grounded justification: metadata-based explanations are designed to be modular, verifiable, and oriented toward revealing the reasons behind model-user interactions.

Within this landscape, the proposed metadata-based approach stands out for its focus not only on technical transparency but also on justifying the influence exerted by LLMs. Transparency alone is not a safeguard against manipulation or misuse [34]; without active mechanisms to detect and prevent dishonest behavior such as deceptive prompting, misaligned outputs, or strategic omission, LLMs risk becoming tools of obfuscation rather than clarity. The originality of this perspective, particularly in its emphasis on proactive accountability, will be further discussed in the concluding section.

VI. Technical Roadmap

This section presents a comprehensive experimental evaluation roadmap. It presents a plan to systematically assess the hypothesis, comprising (i) implementation strategy, (ii) evaluation metrics, and (iii) controlled experimental settings for empirical evaluation.

A. Implementation

To enable metadata-driven LLMs' explainability, we propose the high-level Algorithm 1 to capture, process, and integrate explanatory metadata into model outputs. This approach is designed to facilitate not only technical transparency but also user-centered justification by transforming internal model signals into human-readable explanations. We support the selective disclosure of information based on the type of explanation, while maintaining the flexibility for future metadata categories, such as user profiling, bias detection, and semantic salience. The algorithm begins by calling the core inference function LLM_Infer (Line 1), producing both the model's textual response and its internal state given the user's input; then it initializes the configuration array of explanation types—informative, justificatory, precautionary, and causal—that determine which modules may later be invoked (Line 2). The ExtractAllMetadata function is invoked (Line 3), which aggregates all relevant internal signals into a metadata object. It then calls GenerateExplanation (Line 4), passing the collected metadata and configuration so that each enabled explanation module can produce its fragment. The next step is to persist the full metadata object in an audit log for traceability and iterative refinement. Finally, Line 6 integrates the original model output with the assembled explanation and returns the combined, human-readable result.

The *ExtractAllMetadata* method gathers usable metadata in a single map used in generating explanations (Lines 8-16). The GenerateExplanation routine then transforms the generated metadata map into a coherent narrative (Lines 17–26). It iterates over each enabled explanation category (i.e., informative, justificatory, etc.) where the user initially inputs the configurations. Each module ingests the map **M** and produces a self-contained fragment that emphasizes its rhetorical purpose: whether to inform (i.e., "you can see attention peaks on key terms"), justify (i.e., "this paraphrase reflects the system constraints"), warn (i.e., "confidence is only 62% here, so interpret cautiously"), or trace cause (i.e., "this emerged because of semantic clustering with your topic keywords"). All of these fragments are then passed to a merging function that concatenates them into a single, unified explanation alongside the original LLM response.

To support the implementation of the proposed framework, we will experiment with both on-premise and cloud-based LLMs accessible via APIs, including models such as GPT-4, Claude, and LLaMA. Additionally, the explainability workflow will be orchestrated using agent-based frameworks like LangChain, CrewAI, and Autogen, enabling modular integration and metadata generation throughout the interaction.

Algorithm 1: Configurable Metadata-Based Explanation Workflow

```
1  Procedure GenerateExplainedLLMResponse(input):
2      (output, state) ← LLM_Infer(input)
3      config ← [informative, justificatory,
         precautionary, causal]
4      metadata ← ExtractAllMetadata(input,
         output, state)
5      explanation ←
         GenerateExplanation(metadata, config)
6      LogMetadata(metadata)
7      return Integrate(output, explanation)
8  Function ExtractAllMetadata(input, output, state):
9      M ← {}
10     M.tokens ← GetSaliencyMaps(state)
11     M.clusters ← ClusterEmbeddings(input,
         output)
12     M.confidence ← ComputeConfidence(state)
13     M.constraints ← FetchActivatedRules(state)
14     M.alternatives ←
         EnumerateAlternatives(output)
15     M.profile ← GetUserProfile(input.user_id)
16     return M
17 Function GenerateExplanation(metadata, config):
18     explanations ← [ ]
19     foreach type in config do
20         if IsModuleEnabled(type) then
21             module ←
                 LoadExplanationModule(type)
22             part ← module(metadata)
23             Append(explanations, part)
24         end
25     end
26     return MergeExplanations(explanations)
```

B. Evaluation

An evaluation protocol will be implemented to validate the utility of our metadata-based explanations. Accordingly, the perceived reliability from a user's perspective will be assessed using a seven-point Likert scale (i.e., "I trust this explanation") given immediately after the task. The effectiveness of influence will be evaluated through a between-subjects A/B study, comparing decision quality across a control group (LLM output only) and an experimental group (augmented with metadata explanations). Finally, we will analyze adaptivity by correlating self-reported expertise levels with metrics such as explanation length and module count, thereby quantifying how well the system personalizes its explanatory depth.

VII. Conclusion

This proposal of a metadata-based explainability approach makes an original contribution to the GenAI explainability debate, shifting the focus from simply explaining how the model works to clarifying the nature of the influence it exerts. LLMs do not merely generate content, but subtly and implicitly modulate preferences and behaviors. For this reason, this work argues that explainability must evolve from a purely technical tool to an ethical and relational resource, aimed at revealing the reasons behind LLMs' influence.

While the use of metadata is not new to XAI, it is typically limited to technical indicators such as confidence scores and

source rankings. This proposal advances a reflexive use of metadata as narrative elements to explain how input is interpreted, which alternatives are discarded, and what signals or uncertainties guide the model's behavior. These explanations move beyond the output description to clarify the mechanisms of user influence.

Ultimately, this perspective advocates for a multidimensional and adaptive form of explainability, which extends beyond standard XAI practices. Despite having different focuses, all types of metadata-based explanations share the goal of revealing the reasons for the influence exerted. The aim is not only to make the models comprehensible, but also accountable, fostering an interaction that respects the normative principles governing human relations. Although the concrete realization of this approach poses significant challenges, it represents a crucial theoretical step toward truly trustworthy AI.

REFERENCES

[1] W. T. Piriyakulkij, V. Kuleshov, and K. Ellis, "Active preference inference using language models and probabilistic reasoning," *arXiv preprint arXiv:2312.12009*, 2023.

[2] Z. Ma, W. Wu, Z. Zheng, Y. Guo, Q. Chen, S. Zhang, and X. Chen, "Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 146–11 150.

[3] M. Klenk, "(online) manipulation: sometimes hidden, always careless," *Review of Social Economy*, vol. 80, no. 1, pp. 85–105, 2022.

[4] ——, "Algorithmic transparency and manipulation," *Philosophy & Technology*, vol. 36, no. 4, p. 79, 2023.

[5] F. Jongepier and M. Klenk, "Online manipulation: Charting the field," in *The Philosophy of Online Manipulation*. Routledge, 2022, pp. 15–48.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] S. Serrano and N. A. Smith, "Is attention interpretable?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2931–2951.

[10] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," *arXiv preprint arXiv:2104.08758*, 2021.

[11] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[12] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale," *Advances in neural information processing systems*, vol. 35, pp. 30 318–30 332, 2022.

[13] X. Ma, G. Fang, and X. Wang, "Llm-pruner: On the structural pruning of large language models," *Advances in neural information processing systems*, vol. 36, pp. 21 702–21 720, 2023.

[14] E. Perez, S. Ringer, K. Lukosiute, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath *et al.*, "Discovering language model behaviors with model-written evaluations," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 13 387–13 434.

[15] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4302–4310.

[16] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," 2020. [Online]. Available: https://arxiv.org/abs/1904.09751

[17] J. Schneider, "Explainable generative ai (genxai): A survey, conceptualization, and research agenda," *Artificial Intelligence Review*, vol. 57, no. 11, p. 289, 2024.

[18] J. Schneider, S. Haag, and L. C. Kruse, "Negotiating with llms: Prompt hacks, skill gaps, and reasoning deficits," in *International Conference on Computer-Human Interaction Research and Applications*. Springer, 2024, pp. 238–259.

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[20] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68 539–68 551, 2023.

[21] K. Yin and G. Neubig, "Interpreting language models with contrastive explanations," *arXiv preprint arXiv:2202.10419*, 2022.

[22] R. Carli, A. Najjar, and D. Calvaresi, "Risk and exposure of xai in persuasion and argumentation: The case of manipulation," in *International workshop on explainable, transparent autonomous agents and multi-agent systems*. Springer, 2022, pp. 204–220.

[23] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger *et al.*, "Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions," *Information Fusion*, vol. 106, p. 102301, 2024.

[24] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities," *Information systems management*, vol. 39, no. 1, pp. 53–63, 2022.

[25] T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell, "Toward transparent ai: A survey on interpreting the inner structures of deep neural networks," in *2023 ieee conference on secure and trustworthy machine learning (satml)*. IEEE, 2023, pp. 464–483.

[26] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh *et al.*, "Taxonomy of risks posed by language models," in *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2022, pp. 214–229.

[27] F. Cabitza, A. Campagner, G. Malgieri, C. Natali, D. Schneeberger, K. Stoeger, and A. Holzinger, "Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable ai," *Expert systems with Applications*, vol. 213, p. 118888, 2023.

[28] Z. Chen, J. Chen, C. YuanYuan, H. Yu, A. Singh, and M. Sra, "Lmexplainer: A knowledge-enhanced explainer for language models," 2023.

[29] A. Creswell and M. Shanahan, "Faithful reasoning using large language models," *arXiv preprint arXiv:2208.14271*, 2022.

[30] H. Wang and K. Shu, "Explainable claim verification via knowledge-grounded reasoning with large language models," *arXiv preprint arXiv:2310.05253*, 2023.

[31] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[32] D. Slack, S. Krishna, H. Lakkaraju, and S. Singh, "Explaining machine learning models with interactive natural language conversations using talktomodel," *Nature Machine Intelligence*, vol. 5, no. 8, pp. 873–883, 2023.

[33] M. Turpin, J. Michael, E. Perez, and S. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," *Advances in Neural Information Processing Systems*, vol. 36, pp. 74 952–74 965, 2023.

[34] S. Chern, Z. Hu, Y. Yang, E. Chern, Y. Guo, J. Jin, B. Wang, and P. Liu, "Behonest: Benchmarking honesty in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.13261