Mitigating Misleadingness in LLM-Generated Natural Language Explanations for Recommender Systems: Ensuring Broad Truthfulness Through Factuality and Faithfulness

Ulysse Maes[1,*], Lien Michiels[1,2] and Annelien Smets[1]

[1] *imec-SMIT, Vrije Universiteit Brussel, Brussels, Belgium*
[2] *University of Antwerp, Antwerp, Belgium*

Abstract

Large Language Models (LLMs) are increasingly used to generate natural language explanations in recommender systems, offering the potential for adaptive, personalized, and interactive communication. However, their ability to produce plausible yet inaccurate justifications poses a risk of misleading users at scale. This paper examines the risk of misleading LLM-generated explanations, focusing on the interplay between truthfulness and persuasiveness. We propose a clear terminology for categorizing misleading explanations, differentiating between factual errors and unfaithful representations of the system's workings. Drawing on research from machine learning, human-computer interaction, and persuasive communication, we identify key challenges in evaluating and mitigating misleading explanations. We outline directions for future research, including the development of robust evaluation metrics and methods for enhancing the factuality and faithfulness of LLM-generated explanations in recommender systems.

Keywords

Explainable recommender systems, large language models, evaluating explanations, trustworthiness

1. Introduction

Large Language Models (LLMs) are increasingly employed to generate explanations across various domains. Their flexible Natural Language Processing (NLP) capabilities allow them to enhance the intelligibility of complex outcomes. This is evident in applications such as elucidating the reasoning behind medical diagnoses for patients [1] and translating technical eXplainable AI (XAI) techniques, like SHapley Additive exPlanations (SHAP) values [2], into layman's terms [3]. Furthermore, LLMs enable novel forms of interaction between a user and a digital system, for example, in conversational recommender systems (CRS) [4]. LLMs present particularly promising avenues for explaining personalized recommendations made by recommender systems (RS). This is largely due to their adaptability to diverse contexts, items, and users. As the core purpose of RS lies in achieving scalable personalization [5], LLMs extend

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy

[*] Corresponding author.

ulysse.jan.l.maes@vub.be (U. Maes); lien.michiels@vub.be (L. Michiels); annelien.smets@vub.be (A. Smets)
0009-0005-6823-0237 (U. Maes); 0000-0003-0152-2460 (L. Michiels); 0000-0003-4771-7159 (A. Smets)

CEUR Workshop Proceedings (CEUR-WS.org)

this capability by enabling the personalization of both content and writing style (including tone of voice and complexity) at scale [6, 7]. However, because of their flexibility, adaptivity and scalability, they come with novel risks. These range from over-personalisation [8] to excessive energy consumption with significant environmental impact [9]. This paper focuses on the risk of LLMs misleading users by generating inaccurate yet persuasive explanations for personalized recommendations. We argue that this phenomenon poses a significant challenge to the trustworthy deployment of LLMs in RS.

Moreover, this issue is becoming increasingly pressing as technical boundaries, for example, cost and latency, to the 'production readiness' of LLM-generated explanations are being overcome [10]. If their potential to mislead users is not addressed, it renders them unreliable from a normative perspective [11], as user studies indicate that explanations have a large effect on the item selections [12]. This concern is further compounded by other risk factors: RS rely heavily on personalization, which significantly amplifies their persuasive power [13]. Additionally, RS often employ complex algorithms which allow room for incompleteness or framing, making it difficult to verify their factuality and faithfulness.

Judging precisely how much of a challenge this phenomenon poses is complicated by the fact that earlier work on misleadingness of LLM-generated explanations is limited and spread out over different fields, including machine learning (ML) [14], human-computer interaction (HCI) [15], and computational language [4]. We contribute to the debate by bringing together insights from these different fields and applying them to the context of RS.

In the remainder of this paper, we start with providing a precise description of the problem, which brings us to our first argument and contribution; establishing a shared terminology for this issue. We present the notion of broad truthfulness, consisting out of factuality and faithfulness, as a necessary condition to avoid misleading explanations. We then relate broad truthfulness to similar terms from the literature and ground the concept in a taxonomy of explainable RS. Finally, we discuss what limited research has been done that touches upon misleadingness, and outline directions for future research based on related literature from different fields.

2. Problem Statement

We start by defining the problem of misleading LLM-generated explanations and its causes. Importantly, we focus specifically on natural language explanations of personalized recommendations –generated by LLMs– aimed at end users of any kind and designed to provide insights into the algorithm's functioning [16, 17, 18]. The explanations hereby answer one or more of the following questions: *WHY are these items recommended?* and *HOW does the algorithm work* [19]. Less frequently, explanations may address questions like *what if*, *why not*, and *how to* [20, 18]. This specific focus enables a targeted analysis of the challenges in generating accurate and trustworthy explanations for personalized natural language recommendations.

We argue that LLM-generated explanations are misleading when they are both *inaccurate* and also *persuasive*. In the remainder of this Section, we describe precisely how, when and why LLM-generated explanations meet these conditions and can thus mislead users. We begin by demonstrating that LLMs frequently generate inaccuracies. Building on this, we argue that such inaccuracies have been shown to effectively persuade users into accepting suboptimal recommendations. Finally, we discuss why misleading explanations are undesirable, examining the issue from multiple perspectives.

2.1. LLMs Frequently Generate Plausible but Inaccurate Texts

While LLMs excel at producing plausible narratives that appear logical and believable, their outputs often contain inaccuracies. These inaccuracies stem from various sources, including hallucinations, bias and limitations in the attention mechanism.

Hallucinations appear as factual or contextual inaccuracies in LLM outputs caused by issues with data, training or inference [21, 22]. These are difficult to solve entirely, since LLMs struggle with recognizing their own knowledge boundaries, particularly in long-tail knowledge, recent information, or copyright-sensitive content [21].

Bias is a related issue, which risks to reinforce echo chambers and marginalize minority voices [23]. Bias can stem from imbalanced training data, but can also emerge during the LLM alignment phase [23]. In the context of recommender systems, one example is the generation of explanations that emphasize stereotypical male or female aspects based on the user's gender [24].

Literature also highlights different *limitations in the attention mechanism*. Firstly, LLMs tend to prioritize user utterances over broader context [4]. This is exemplified in the second example depicted in Table 1, where a content-based movie CRS misidentifies key drivers due to its bias towards a user's latest input. Secondly, earlier context, such as a system prompt, previous explanations, or information from a knowledge base, may be overlooked or forgotten [25].

2.2. Inaccurate Explanations are Likely to Mislead Users

Inaccurate explanations become misleading when they succeed in persuading the user to accept a flawed recommendation. LLMs frequently employ confident and persuasive language, by using linguistic techniques such as confidence manipulation, appeals to authority, and selective presentation of evidence [26]. For example, Danry et al. [13] conducted an online user study demonstrating that LLM-generated inaccurate explanations can be more persuasive than accurate and honest ones. This effect is attributed to an LLM's capability to produce logically coherent justifications for incorrect information. While persuasion is not exclusive to language models, LLMs excel at targeted personalization—an effective persuasion technique [13]. Moreover, LLMs enable personalized persuasion on a larger scale and at a lower cost [27, 28, 13]. Further research supports the effectiveness of LLM-generated texts, showing it can surpass human-written content in persuasive [4, 29, 30] and misleading [31] communication. Sadeghi et al. [15] highlight that persuasive explanations can foster unwarranted trust in the system, increasing users' confidence in the accuracy of incorrect predictions. Hence, misleading explanations can effectively persuade users to accept recommendations, even if those recommendations are not genuinely in their best interests [32, 33].

2.3. Misleading Explanations Lead to Undesirable Consequences

The persuasive power of misleading explanations presents a twofold danger. Firstly, it can foster unfounded trust and an illusion of transparency regarding the system's operation. In this way, large groups of users can be manipulated into consuming certain items under false pretenses. Secondly, it risks exacerbating existing power imbalances between users and platform owners, further tilting the scales in favor of the latter. Given that persuasion is often an important objective in RS [34], ensuring the truthfulness of explanations becomes paramount to avoid misleadingness. This is crucial from normative, practical, and regulatory perspectives.

Normatively, explanations in RS must be faithful, accurately reflecting underlying processes to ensure fairness [35]. This requires transparent communication of algorithmic objectives [36, 37]. Even well-intentioned efforts like taste-broadening may be rejected if seen as manipulative or imposing a contested notion of "fairness" [37].

Practically, while deceptive explanations might offer short-term benefits, they risk long-term damage to the credibility of the system and the entities that use them. Conversely, when users genuinely comprehend how a system operates, it leads to greater satisfaction and trust [33, 37, 34, 38, 39]. Specifically for CRS, more credible explanations lead to a higher-quality conversation context, which improves accuracy and persuasion [4].

From a *regulatory* standpoint, various guidelines and acts emphasize the importance of true transparency in AI systems. The European ethics guidelines for trustworthy AI [1], DSA [2], the EU AI Act [40] and the USA executive order on safe, secure and trustworthy AI [3] all highlight the need for transparency and human oversight in the development and deployment of AI systems.

3. Building a Shared Terminology

3.1. Defining Misleadingness

In the problem statement, we defined "misleadingness" as the persuasive presentation of plausible yet inaccurate information. However, this definition leaves the concept of "accuracy" ambiguous. This lack of a precise definition has two negative consequences. Firstly, it hinders collaboration among researchers and impedes the comparability of research findings. Secondly, as Lipton highlights, such "conceptual murkiness provides a real opportunity to mislead" without accountability for platform owners or developers [41].

To address this ambiguity, we propose a formal definition of "broad truthfulness", building upon the work of Evans et al. [42, 27]. This clarifies the term "inaccurate explanations" as referring to "explanations that do not comply with the requirements for broad truthfulness". Our definition of "broad truthfulness" encompasses two key dimensions: *factuality* and *faithfulness*. *Factuality* specifically emphasizes the avoidance of falsehoods, aligning with what Evans et al. term "narrow truthfulness". Related terms found in the literature include "correctness", "veracity", and "factfulness".

In contrast, "broad truthfulness" extends beyond factuality to include the avoidance of both outright falsehoods and an inaccurate framing which instills an incorrect mental model with the user. This concept emphasizes informativeness and proper representation, ensuring explanations

[1] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[2] Article 27 of the regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Mark et For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277.

[3] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

are not only factually accurate but also contextually relevant and sufficiently comprehensive. We refer to this second component of "broad truthfulness" as *faithfulness* [43]: the congruence between statements and the underlying beliefs or processes. *Faithfulness* can be further broken down into "sincerity", ensuring that explanations are consistent with the recommended actions, and "transparency", providing real insight into the system's underlying mechanisms [44]. Faithfulness is however not equal to exhaustiveness. For example, explanations can simplify the reasoning process by leaving out non-essential information.

3.2. Evaluating Misleadingness

To evaluate the misleadingness of explanations in a holistic manner, we adopt the four perspectives identified by Ge et al. [45] in their 2022 survey on explainable recommender systems. These perspectives offer a comprehensive framework for analyzing the quality of explanations, ensuring a nuanced assessment from multiple angles. Furthermore, these perspectives can also be applied to assess the truthfulness of explanations. In this Section, we present these perspectives and discuss their relevance to the concepts of factuality and faithfulness.

- **Explanation Method**: Explanations can be divided into *model-intrinsic* ones, which are linked to the model's internal framework, and *model-agnostic* ones, which are independent of the model's architecture [43]. The strategy for maintaining explanation factuality varies with the RS type. With white-box models, whose internal processes are transparent, model-intrinsic explainers utilize the system's logic. This enables factual explanation but may hinder system performance [43]. In contrast, black-box RS require model-agnostic explanation methods, also called "justifications". "Justifications" provide plausible narratives based on the output, independent of the RS's internal workings [46]. While they offer valuable insights, it's crucial to remember that they may contain factual errors [26]. Therefore, they should faithfully communicate any uncertainties or approximations. When an LLM is used for recommendation elicitation itself, it is debatable whether true model-intrinsic explanations can be generated, although certain prompting methods might provide some level of transparency, as discussed in Section 4.2.1, "Prompt-based Mitigation".
- **Explanation Scope**: The model's decision-making rationale can be explained from a *global perspective*, providing users with insights into the entire information flow and decision processes of the model. Conversely, a *local explanation scope* focuses specifically on the reasoning behind individual recommended items. Typically, natural language explanations align more closely with a local explanation scope by focusing on why a particular user got recommended a specific set of items. A local scope, while useful for explaining individual recommendations, can sometimes omit crucial details about the system's overall functioning. This selective focus can lead to a myopic view of the situation and potentially mislead users about the broader context. To ensure faithful explanations, it is essential to clearly indicate the extent to which the explanation generalizes to the entire system. If such generalizability cannot be guaranteed, it should not be implied. This sincerity helps users to accurately interpret the explanation within the broader context of the RS's operation.

Algorithm	Misleading Explanation	Inaccuracy	Truthful Explanation
Collaborative Filtering	"We recommend *The Bourne Identity* because it is an action movie with a spy protagonist similar to your preference for *James Bond* and *Mission Impossible*."	Faithful	"We recommend *The Bourne Identity* because users who liked action movies with a spy protagonist like *James Bond* and *Mission Impossible* liked this too."
Content-Based	"Based on your preference for *Casino Royale*, I recommend *The Bourne Identity*, because it is also in Paris."	Factual	"Based on your preference for *Casino Royale*, I recommend *The Bourne Identity*, because it's also an action-packed spy movie with A-list actors."

Table 1
Two hypothetical examples illustrating the need for both factuality and faithfulness in explanations provided by recommender systems.

- **Affected User**: This term was adapted from "benefited user" from Ge et al. [45]. We prefer to use the term "affected user" as people can be affected by an explanation without benefiting from it. Individuals impacted by explanations include the system owner, the user, and the model designers, as highlighted by Chen et al. [47]. Other affected users may be present, such as content providers on multi-sided platforms. Different users pursue different "explanation goals" [48], necessitating distinct explanations [49]. System owners, developers and item providers may have incentives to persuade and hence may be tempted to not fully consider the potential misleadingness of an explanation, or in exceptional cases, resort to outright deception. Moreover, users' perceptions of explanations can vary significantly and may not align with the objective characteristics of the system. Knijnenburg and Willemsen present a framework to separate these Objective System Aspects (OSA) from the Subjective System Aspects (SSA) [17]. By evaluating SSA with relevant users before deployment to production, system developers can develop user models that consider potential conflicts and synergies among stakeholders.
- **Explanation Style**: This dimension reflects the system's assumed inputs and reasoning. The explanation style can be *case-based*, *collaborative-based*, *content-based*, *conversational*, or *demographic-based*. While the explanation style may align with the algorithm, a recommendation can also be framed in a different one, for instance to make the system look more advanced, or more privacy-friendly. Because of this potential to mislead, earlier research highlights the importance of using an explanation style that matches the RS to ensure faithfulness [50, 44]. To illustrate this, Table 1 presents a justification which is not faithful by using a content-based explanation style, while a collaborative filtering algorithm has been used.

4. Future Research Directions

4.1. Evaluation Metrics

Different experimental designs to evaluate the quality and effects of explanations in RS can be categorized in either offline, or online experiments. While offline experiments are generally more accessible, they provide limited insights into the SSA [17]. In contrast, online experiments involve real users, which makes them generally more expensive to carry out, but provide a more realistic assessment by enabling qualitative evaluation [47]. To assess misleadingness of explanations in a RS, we need to measure all three components, namely, factuality, faithfulness and persuasiveness.

Misleadingness in explanations arises from a discrepancy between the OSA and SSA. Therefore, evaluating misleadingness requires a dual approach. Firstly, we need to objectively measure the factuality and faithfulness of the explanation's language, ensuring it accurately reflects the OSA. Secondly, we must assess the impact of explanations on user behavior, specifically examining the change in acceptance of recommendations attributable to the explanations. This helps gauge the persuasive power of the explanations and their potential to influence user choices.

4.1.1. Offline Evaluation

Several existing offline evaluation methods are proposed for the assessment of **factuality**. One approach is to use "perplexity", which measures how unlikely it is for an LLM to produce a given string, with lower perplexity indicating more reasonable assumptions [3]. However, the reliability of perplexity has been criticized [3, 51]. Another option is to calculate the Mean Explainability Precision score, which returns 1 if the entire model is explained, though it depends on a clear definition of interpretability [52]. Future work could explore a more direct measure of factuality, for example, in the form of a fact-checking pipeline which identifies statements and assigns a score indicating the proportion of truthful statements.

While offline evaluation of **faithfulness** is inherently limited by the subjective nature of misleadingness, some proxies have been proposed. For example, Xu et al. [43] suggest evaluating faithfulness by measuring the overlap of key inputs between the decision model and the explanation model. This involves focusing on the shared importance of input features related to the output, and is particularly relevant when using a white-box trained explanation model on a black-box RS. However, the concept of calculating the semantic similarity between a natural language explanation and the RS output in general offers a promising direction for future research. Another option is to use an "Explainability Score" that measures the number of user interactions in the training dataset that support the explanation [52]. The same research also proposes "Model Fidelity" as a metric for the alignment between an explainer model and the explained RS algorithm [52].

Measuring **persuasion** offline is challenging due to its inherent psychological nature. However, recognizing that subtle changes in wording can significantly impact persuasiveness [12], future work could involve developing a comparable multilingual benchmark to identify unfaithful, persuasive, or deceptive language patterns from user studies. These insights could then be applied in an offline manner to assess the persuasive potential of explanations.

4.1.2. Online Evaluation

While offline methods offer valuable insights, truly understanding the user experience with explanations in the dynamic and personalized context of RS necessitates user experiments [17]. Online experiments allow for measuring perceived values, which is essential for concepts where objective and subjective aspects may diverge.

Prior research has explored **user perceptions** of explanations using methods like Likert scales and free-form interviews. To assess the dimensions of misleadingness, researchers employ a mix of direct questions (i.e. "Do you believe this explanation is accurate?") and behavioral observations (i.e. measuring time spent reading the explanation or item selection). Note that perceived transparency and trust may have a complicated relationship: Increased transparency may, for example, foster trust in the system if the user feels less manipulated, but it may also expose flaws in the system making the user lose trust. Additionally, there might be differences in actual **behavior** versus self-reported behavior. For example, perceived persuasiveness can be related to actual acceptance of recommendation, which can be approximated with common engagement metrics such as CTR. However, to measure the specific impact of recommendations on item selection, future work could explore a more targeted metric to define an “acceptance ratio”, which measures the uplift or downlift in item selection observed after presentation of the explanation.

4.2. Methods to Improve Factuality and Faithfulness

Based on our survey of the literature on explanation generation in the domains of LLMs, HCI and persuasive recommendation, we identify three main categories of mitigation strategies: *prompting*, *interface* and *model-based.* Table 2 provides a schematic overview of the mitigation strategies and the areas – factuality or faithfulness – they address.

4.2.1. Prompt-based Mitigation

Prompt-based mitigation methods involve strategies designed to instruct LLMs in a way that improves the reliability of their outputs.

A first family of methods focuses on knowledge enrichment by providing relevant information to the context. This information can come from two different sources. A first option is to *leverage classical XAI methods*, for example, SHAP [2], perturbation tests and knowledge distillation. These methods produce factual, yet possibly complex explanations which can be then simplified in natural language [3, 43, 53]. A preliminary survey by Mavrepis et al. [54] reports positive effects of on both end-users as AI experts in describing classical XAI techniques in a more effective and understandable way, with 75% of participants preferring their “x-[plAIn]” prompting approach.

A second option is to *provide external knowledge* on users, items, or interactions using knowledge graphs or techniques like Retrieval-Augmented Generation (RAG) with source attribution [55]. For example, the Parametric Knowledge Guiding (PKG) framework integrates domain-specific knowledge to enhance LLM factuality [56, 21]. Future work could explore how to optimally leverage injected knowledge, as LLMs tend to prefer internal beliefs, even when they contradict external sources [3, 57, 58].

Another prompting approach is to validate outputs after generation. This includes *self-reflection* methods, such as internal lie detection, identification of knowledge boundaries, and using probing techniques to detect falsehoods [21, 59, 60]. In contrast, *external validation* methods utilize another LLM to identify unfaithful language, block questionable content, or generate disclaimers [27]. An external LLM can also decompose explanations into atomic statements and verify each one using an external truth-checking API [21]. Zhang et al. [61] found that leveraging external LLMs to assess the quality of textual explanations is a scalable method that aligns well with human feedback, and that an ensemble of LLMs further enhances both the accuracy and stability of the evaluation.

Another family of prompting methods focuses on transparency. Direct *transparent prompting techniques*, such as chain-of-thought (CoT), can enhance the LLM's accuracy and provide transparency about the reasoning process to users [62, 63, 44]. Furumai et al. [64] present a small-scale simulated user experiment following an approach that combines atomic splitting of statements and fact-checking them using a CoT prompt, which they found to improve both persuasiveness and factuality. However, these approaches can still yield unfaithful results and tend to be slower and more expensive due to increased token consumption [65]. Next to explicit prompting, *confidence* in the LLM's output—defined as the average next-token probability—offers an indirect method for estimating how confident the model is about the explanation [21, 66, 67]. Future work could explore if confidence can be reliably improved by setting the model's temperature to lower values. For example, the data analytics company Tickr reported a trade-off between higher temperature and factually correct responses [68]. The LLM's confidence is not to be confused with the RS's confidence in the predicted items, which could also be leveraged to calibrate the tone of explanations, rendering less persuasive explanations for uncertain predictions [15].

Finally, LLMs may become targets for manipulation. For example, item providers could post reviews containing "jailbreaks" to influence the output of LLM-based review summarizers [45]. System prompts should therefore contain sufficient guard rails to ensure *robustness against adversarial attacks*.

4.2.2. Interface-based Mitigation

Interface-based mitigation methods primarily aim to enhance the explanation goals of "transparency" and "scrutability" by making elements visible and/or editable by the user. An often-suggested method is to *label explanations as AI-generated*. However, user experiments show that such labeling currently consistently reduces trust [69, 70]. For example, two-large scale user experiments conducted by Wittenberg et al. [71] on AI-generated misleading information show that adding labels decreased people's likelihood of believing the content, although the impact of different labels varied significantly. An alternative is to take a further step and *make the prompt visible*, providing transparency to users interested in understanding the process behind the output. Here, future work could investigate the right balance between the availability of this information while avoiding information overload [19]. Additionally, we can draw inspiration from the field of *scrutable interfaces* [72, 73]. Future research could explore natural language interfaces that give users more control over the assumptions the system makes about them.

Implementation	Method	Factuality	Faithfulness	Sources
Prompting	Leverage classical XAI methods	X	X	[3]
	Provide external knowledge	X		[3, 57, 58, 56]
	Self-reflection	X		[4, 21]
	External validation	X	X	[27]
	Transparent prompting	X	X	[44, 63, 62, 65]
	Leverage confidence	X	X	[15]
	Robustness against adversarial attacks	X	X	[45]
Interface	Label as AI-generated		X	[71, 69, 70]
	Make prompt visible		X	[19]
	Scrutable interfaces	X		[72, 73]
Model	Boost interpretable features		X	[74]
	Fine-tune on explainable RS settings		X	[75, 21, 44]
	Alignment with desired behavior	X	X	[77, 78]

Table 2
Overview of possible methods to ensure factual and faithful explanations

4.2.3. Model-based Mitigation

Thirdly, *model-based* mitigation approaches involve developing specialized language models, which typically requires substantial data and computing resources. One early research direction is the *extraction and boosting of interpretable features*, such as honesty, which might help align the internal state of an LLM more closely with the system's goals [74]. Another model-based research direction involves *fine-tuning* the model on carefully selected explainable RS examples, enabling the model to generate explanations in a more appropriate manner while recognizing its own limitations and guardrails [75, 21, 44]. In addition to fine-tuning, *alignment with desired behavior* can be achieved through techniques such as Reinforcement Learning with Human Feedback (RLHF), or using a small set of demonstrations [76], which aims to bring the outputs in line with normative goals [77]. However, more research on the reliability of alignment techniques is needed, as recent studies show that LLMs can fake alignment [78].

5. Conclusion

In this paper, we examined the emerging issue of misleading explanations generated by Large Language Models (LLMs) in recommender systems. We highlighted the potential for seemingly plausible yet inaccurate explanations to mislead users, particularly given the persuasive capabilities of LLMs. By proposing a clear terminology and drawing on research from various fields, we aimed to provide a comprehensive understanding of the problem and identify key areas for future exploration. Future research should develop robust evaluation metrics and mitigation strategies to ensure that persuasive LLM-generated explanations are both factual and faithful, fostering transparency and trustworthiness in recommender systems.

Acknowledgements

This research was supported by the Research Foundation Flanders (FWO) under grant number 1158225N.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT-4o) and Google Gemini Pro in order to paraphrase and reword, improve writing style as well as for abstract drafting. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Zhang, J. Li, Y. Wang, Y. Song, Integrating automated knowledge extraction with large language models for explainable medical decision-making, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2023, pp. 1710–1717. doi:10.1109/BIBM58861.2023.10385557.

[2] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[3] T. Ichmoukhamedov, J. Hinns, D. Martens, How good is my story? towards quantitative metrics for evaluating llm-generated xai narratives, 2024. URL: https://arxiv.org/abs/2412.10220. arXiv:2412.10220.

[4] P. Qin, C. Huang, Y. Deng, W. Lei, T.-S. Chua, Beyond persuasion: Towards conversational recommender system with credible explanations, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 4264–4282. URL: https://aclanthology.org/2024.findings-emnlp.247/. doi:10.18653/v1/2024.findings-emnlp.247.

[5] G. Adomavicius, Z. Huang, A. Tuzhilin, Personalization and Recommender Systems, INFORMS, 2008, pp. 55–107. URL: http://dx.doi.org/10.1287/educ.1080.0044. doi:10.1287/educ.1080.0044.

[6] C. Li, M. Zhang, Q. Mei, Y. Wang, S. A. Hombaiah, Y. Liang, M. Bendersky, Teach llms to personalize – an approach inspired by writing education, 2023. URL: https://arxiv.org/abs/2308.07968. arXiv:2308.07968.

[7] E. Meguellati, L. Han, A. Bernstein, S. Sadiq, G. Demartini, How good are llms in generating personalized advertisements?, in: Companion Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 826–829. URL: https://doi.org/10.1145/3589335.3651520. doi:10.1145/3589335.3651520.

[8] T. Lazovich, Filter bubbles and affective polarization in user-personalized large language model outputs, in: J. Antorán, A. Blaas, K. Buchanan, F. Feng, V. Fortuin, S. Ghalebikesabi,

A. Kriegler, I. Mason, D. Rohde, F. J. R. Ruiz, T. Uelwer, Y. Xie, R. Yang (Eds.), Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops, volume 239 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 29–37. URL: https://proceedings.mlr.press/v239/lazovich23a.html.

[9] P. Jiang, C. Sonne, W. Li, F. You, S. You, Preventing the immense increase in the life-cycle energy and carbon footprints of llm-powered intelligent chatbots, Engineering 40 (2024) 202–210. URL: https://www.sciencedirect.com/science/article/pii/S2095809924002315. doi:https://doi.org/10.1016/j.eng.2024.04.002.

[10] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, J. Zhang, Chat-rec: Towards interactive and explainable llms-augmented recommender system, 2023. URL: https://arxiv.org/abs/2303.14524. arXiv:2303.14524.

[11] T. Ma, Y. Cheng, H. Zhu, H. Xiong, Large language models are not stable recommender systems, 2023. URL: https://arxiv.org/abs/2312.15746. arXiv:2312.15746.

[12] K. Balog, F. Radlinski, A. Petrov, Measuring the impact of explanation bias: A study of natural language justifications for recommender systems, in: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, ACM, 2023, p. 1–8. URL: http://dx.doi.org/10.1145/3544549.3585748. doi:10.1145/3544549.3585748.

[13] V. Danry, P. Pataranutaporn, M. Groh, Z. Epstein, P. Maes, Deceptive ai systems that give explanations are more convincing than honest ai systems and can amplify belief in misinformation, 2024. URL: https://arxiv.org/abs/2408.00024. arXiv:2408.00024.

[14] K. Lee, S. Ram, Explainable deep learning for false information identification: An argumentation theory approach, Information Systems Research 35 (2024) 890â€"907. URL: http://dx.doi.org/10.1287/isre.2020.0097. doi:10.1287/isre.2020.0097.

[15] M. Sadeghi, D. Pöttgen, P. Ebel, A. Vogelsang, Explaining the Unexplainable: The Impact of Misleading Explanations on Trust in Unreliable Predictions for Hardly Assessable Tasks, in: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 36–46. URL: https://dl.acm.org/doi/10.1145/3627043.3659573. doi:10.1145/3627043.3659573, [Online; accessed 2024-10-28].

[16] W. Fan, X. Zhao, X. Chen, J. Su, J. Gao, L. Wang, Q. Liu, Y. Wang, H. Xu, L. Chen, Q. Li, A comprehensive survey on trustworthy recommender systems, 2022. URL: https://arxiv.org/abs/2209.10117. arXiv:2209.10117.

[17] B. P. Knijnenburg, M. C. Willemsen, Evaluating Recommender Systems with User Experiments, Springer US, Boston, MA, 2015, pp. 309–352. URL: https://doi.org/10.1007/978-1-4899-7637-6_9. doi:10.1007/978-1-4899-7637-6_9.

[18] K. Zhang, Q. Cao, F. Sun, Y. Wu, S. Tao, H. Shen, X. Cheng, Robust recommender system: A survey and future directions, 2023. URL: https://arxiv.org/abs/2309.02057. arXiv:2309.02057.

[19] A. Grün, X. Neufeld, Transparently serving the public: Enhancing public service media values through exploration, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1045–1048. URL: https://doi.org/10.1145/3604915.3610243. doi:10.1145/3604915.3610243.

[20] M. G. Qurat Ul Ain, Mohamed Amine Chatti, S. Joarder, A multi-dimensional conceptu-

alization framework for personalized explanations in recommender systems, in: Joint Proceedings of the ACM IUI Workshops 2022, March 2022' , url='https://ceur-ws.org/Vol-3124/paper2.pdf, 2022.

[21] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. (2024). URL: https://dl.acm.org/doi/10.1145/3703155. doi:10.1145/3703155, just Accepted.

[22] Y. Lin, D. Ghose, D. Coates, J. You, Towards Sustainability of Large Language Models for Recommender Systems, Proceedings of the 18th ACM Conference on Recommender Systems (2024).

[23] Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, S. S. Liu, Bias in Large Language Models: Origin, Evaluation, and Mitigation, 2024. URL: http://arxiv.org/abs/2411.10915. doi:10.48550/arXiv.2411.10915, arXiv:2411.10915 [cs].

[24] Y. Wan, G. Pu, J. Sun, A. Garimella, K.-W. Chang, N. Peng, "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters, 2023. URL: https://arxiv.org/abs/2310.09219. arXiv:2310.09219.

[25] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, D. Yu, Longmemeval: Benchmarking chat assistants on long-term interactive memory, 2024. URL: https://arxiv.org/abs/2410.10813. arXiv:2410.10813.

[26] R. Ajwani, S. R. Javaji, F. Rudzicz, Z. Zhu, Llm-generated black-box explanations can be adversarially helpful, 2024. URL: https://arxiv.org/abs/2405.06800. arXiv:2405.06800.

[27] M. Burtell, T. Woodside, Artificial influence: An analysis of ai-driven persuasion, 2023. URL: https://arxiv.org/abs/2303.08721. arXiv:2303.08721.

[28] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, User Modeling and User-Adapted Interaction 27 (2017) 393–444. doi:10.1007/s11257-017-9195-0.

[29] A. Rogiers, S. Noels, M. Buyl, T. D. Bie, Persuasion with large language models: a survey, 2024. URL: https://arxiv.org/abs/2411.06837. arXiv:2411.06837.

[30] T. H. Costello, G. Pennycook, D. G. Rand, Durably reducing conspiracy beliefs through dialogues with AI, Science 385 (2024) eadq1814. doi:10.1126/science.adq1814, publisher: American Association for the Advancement of Science.

[31] C. Chen, K. Shu, Can LLM-Generated Misinformation Be Detected?, 2024. URL: http://arxiv.org/abs/2309.13788. doi:10.48550/arXiv.2309.13788, arXiv:2309.13788 [cs].

[32] M. Bilgic, R. J. Mooney, Explaining recommendations: Satisfaction vs. promotion, in: Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces, San Diego, CA, 2005. URL: http://www.cs.utexas.edu/users/ai-lab?bilgic:iui-bp05.

[33] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM conference on Computer supported cooperative work, CSCW '00, Association for Computing Machinery, New York, NY, USA, 2000, pp. 241–250. URL: https://dl.acm.org/doi/10.1145/358916.358995. doi:10.1145/358916.358995, [Online; accessed 2024-12-23].

[34] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, 2007. doi:10.1109/ICDEW.2007.4401070.

[35] R. Visser, T. M. Peters, I. Scharlau, B. Hammer, Trust, distrust, and appropriate reliance in (x)ai: a survey of empirical evaluation of user trust, 2023. URL: https://arxiv.org/abs/2312.02034. arXiv:2312.02034.

[36] M. Elahi, D. Jannach, L. Skjærven, E. Knudsen, H. Sjøvaag, K. Tolonen, Holmstad, I. Pipkin, E. Throndsen, A. Stenbom, E. Fiskerud, A. Oesch, L. Vredenberg, C. Trattner, Towards responsible media recommendation, AI and Ethics 2 (2022). doi:10.1007/s43681-021-00107-7.

[37] N. Sonboli, J. J. Smith, F. Cabral Berenfus, R. Burke, C. Fiesler, Fairness and transparency in recommendation: The users' perspective, in: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 274–279. URL: https://doi.org/10.1145/3450613.3456835. doi:10.1145/3450613.3456835.

[38] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, J. Riedl, Is seeing believing? how recommender system interfaces affect users' opinions, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '03, Association for Computing Machinery, New York, NY, USA, 2003, pp. 585–592. URL: https://doi.org/10.1145/642611.642713. doi:10.1145/642611.642713.

[39] J. Koenemann, N. J. Belkin, A case for interaction: a study of interactive information retrieval behavior and effectiveness, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '96, Association for Computing Machinery, New York, NY, USA, 1996, pp. 205–212. URL: https://dl.acm.org/doi/10.1145/238386.238487. doi:10.1145/238386.238487.

[40] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, E. Gomez, The role of explainable AI in the context of the AI Act, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1139–1150. URL: https://dl.acm.org/doi/10.1145/3593013.3594069. doi:10.1145/3593013.3594069.

[41] https://www.facebook.com/48576411181, It's Too Easy to Hide Bias in Deep-Learning Systems — spectrum.ieee.org, https://spectrum.ieee.org/its-too-easy-to-hide-bias-in-deeplearning-systems, ???? [Accessed 09-01-2025].

[42] O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, W. Saunders, Truthful AI: Developing and governing AI that does not lie, 2021. URL: http://arxiv.org/abs/2110.06674. doi:10.48550/arXiv.2110.06674, arXiv:2110.06674 [cs].

[43] Z. Xu, H. Zeng, J. Tan, Z. Fu, Y. Zhang, Q. Ai, A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability, ACM Trans. Inf. Syst. 42 (2023) 29:1–29:29. URL: https://dl.acm.org/doi/10.1145/3605357. doi:10.1145/3605357.

[44] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang, K. Zheng, D. Lian, E. Chen, When large language models meet personalization: perspectives of challenges and opportunities, World Wide Web 27 (2024) 42. doi:10.1007/s11280-024-01276-1.

[45] Y. Ge, S. Liu, Z. Fu, J. Tan, Z. Li, S. Xu, Y. Li, Y. Xian, Y. Zhang, A survey on trustworthy recommender systems, 2022. URL: https://arxiv.org/abs/2207.12515v1. doi:10.48550/

arXiv.2207.12515.

[46] M. Guesmi, M. A. Chatti, S. Joarder, Q. U. Ain, C. Siepmann, H. Ghanbarzadeh, R. Alatrash, Justification vs. transparency: Why and how visual explanations in a scientific literature recommender system, Information (Switzerland) 14 (2023) 23. URL: https://arxiv.org/abs/2305.17034v1. doi:10.3390/info14070401.

[47] X. Chen, Y. Zhang, J.-R. Wen, Measuring "Why" in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation, 2022. URL: http://arxiv.org/abs/2202.06466. doi:10.48550/arXiv.2202.06466, arXiv:2202.06466 [cs].

[48] N. Tintarev, J. Masthoff, Explaining Recommendations: Design and Evaluation, Springer US, Boston, MA, 2015, pp. 353–382. URL: https://link.springer.com/10.1007/978-1-4899-7637-6_10, dOI: 10.1007/978-1-4899-7637-6_10.

[49] D. Gunning, E. Vorm, J. Y. Wang, M. Turek, Darpa's explainable ai (xai) program: A retrospective, Applied AI Letters 2 (2021) e61. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61. doi:https://doi.org/10.1002/ail2.61. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.61.

[50] L. Dong, S. Huang, F. Wei, M. Lapata, M. Zhou, K. Xu, Learning to generate product reviews from attributes, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 623–632. URL: https://aclanthology.org/E17-1059/.

[51] Y. Wang, J. Deng, A. Sun, X. Meng, Perplexity from plm is unreliable for evaluating text quality, 2023. URL: https://arxiv.org/abs/2210.05892. arXiv:2210.05892.

[52] L. Coba, R. Confalonieri, M. Zanker, RecoXplainer: A Library for Development and Offline Evaluation of Explainable Recommender Systems, IEEE Computational Intelligence Magazine 17 (2022) 46–58. URL: https://ieeexplore.ieee.org/abstract/document/9679765?casa_token=mSw5iHaKDAcAAAAA:CrhhgJa2VfEpnl6U81OOuVHjDV__mGdEbg68w71OrPE-wGUlU7cXYkpcljMYpignBG7mJSm-e2c. doi:10.1109/MCI.2021.3129958, conference Name: IEEE Computational Intelligence Magazine.

[53] O. Barkan, V. Bogina, L. Gurevitch, Y. Asher, N. Koenigstein, A Counterfactual Framework for Learning and Evaluating Explanations for Recommender Systems, in: Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 3723–3733. URL: https://dl.acm.org/doi/10.1145/3589334.3645560. doi:10.1145/3589334.3645560.

[54] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, D. Kyriazis, Xai for all: Can large language models simplify explainable ai?, 2024. URL: https://arxiv.org/abs/2401.13110. arXiv:2401.13110.

[55] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. URL: https://arxiv.org/abs/2312.10997. arXiv:2312.10997.

[56] N. Varshney, W. Yao, H. Zhang, J. Chen, D. Yu, A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation, 2023. URL: https://arxiv.org/abs/2307.03987. arXiv:2307.03987.

[57] J. Xie, K. Zhang, J. Chen, R. Lou, Y. Su, Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts, in: The Twelfth International

Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=auKAUJZMO6.

[58] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, S. Singh, Entity-based knowledge conflicts in question answering, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7052–7063. URL: https://aclanthology.org/2021.emnlp-main.565/. doi:10.18653/v1/2021.emnlp-main.565.

[59] A. Azaria, T. Mitchell, The internal state of an llm knows when it's lying, 2023. URL: https://arxiv.org/abs/2304.13734. arXiv:2304.13734.

[60] C. Burns, H. Ye, D. Klein, J. Steinhardt, Discovering latent knowledge in language models without supervision, 2024. URL: https://arxiv.org/abs/2212.03827. arXiv:2212.03827.

[61] X. Zhang, Y. Li, J. Wang, B. Sun, W. Ma, P. Sun, M. Zhang, Large language models as evaluators for recommendation explanations, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 33–42. URL: https://doi.org/10.1145/3640457.3688075. doi:10.1145/3640457.3688075.

[62] Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.-G. Lou, W. Chen, Making language models better reasoners with step-aware verifier, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5315–5333. URL: https://aclanthology.org/2023.acl-long.291/. doi:10.18653/v1/2023.acl-long.291.

[63] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[64] K. Furumai, R. Legaspi, J. Vizcarra, Y. Yamazaki, Y. Nishimura, S. J. Semnani, K. Ikeda, W. Shi, M. S. Lam, Zero-shot persuasive chatbots with llm-generated strategies and information retrieval, 2024. URL: https://arxiv.org/abs/2407.03585. arXiv:2407.03585.

[65] M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL: https://arxiv.org/abs/2305.04388. arXiv:2305.04388.

[66] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. Glass, P. He, Dola: Decoding by contrasting layers improves factuality in large language models, 2024. URL: https://arxiv.org/abs/2309.03883. arXiv:2309.03883.

[67] N. Dziri, A. Madotto, O. Zaïane, A. J. Bose, Neural path hunter: Reducing hallucination in dialogue systems via path grounding, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 2197–2214. URL: https://aclanthology.org/2021.emnlp-main.168/. doi:10.18653/v1/2021.emnlp-main.168.

[68] The Impact of Temperature on the Performance of Large Language Model Systems and Business Applications — tickr.com, https://www.tickr.com/blog/posts/impact-of-temperature-on-llms/, ???? [Accessed 28-01-2025].

[69] S. Altay, F. Gilardi, People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation, PNAS Nexus 3 (2024) pgae403. URL: https://doi.org/10.1093/pnasnexus/pgae403. doi:10.1093/pnasnexus/pgae403.

[70] E. Karinshak, S. X. Liu, J. S. Park, J. T. Hancock, Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–29. URL: https://dl.acm.org/doi/10.1145/3579592. doi:10.1145/3579592.

[71] C. Wittenberg, Z. Epstein, A. J. Berinsky, D. G. Rand, Labeling AI-Generated Content: Promises, Perils, and Future Directions, An MIT Exploration of Generative AI (2024). URL: https://mit-genai.pubpub.org/pub/hu71se89/release/1. doi:10.21428/e4baedd9.0319e3a6, publisher: MIT.

[72] P. Pu, L. Chen, Trust-inspiring explanation interfaces for recommender systems, Knowledge-Based Systems 20 (2007) 542–556. URL: https://www.sciencedirect.com/science/article/pii/S0950705107000445. doi:https://doi.org/10.1016/j.knosys.2007.04.004, special Issue On Intelligent User Interfaces.

[73] J. Kay, B. Kummerfeld, Scrutability, user control and privacy for distributed personalization, ResearchGate (2006).

[74] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, T. Henighan, Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, Transformer Circuits Thread (2024). URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[75] J. Wei, D. Huang, Y. Lu, D. Zhou, Q. V. Le, Simple synthetic data reduces sycophancy in large language models, 2024. URL: http://arxiv.org/abs/2308.03958. doi:10.48550/arXiv.2308.03958, arXiv:2308.03958 [cs].

[76] O. Shaikh, M. Lam, J. Hejna, Y. Shao, M. Bernstein, D. Yang, Show, don't tell: Aligning language models with demonstrated feedback, 2024. URL: https://arxiv.org/abs/2406.00888. arXiv:2406.00888.

[77] Z. Wang, B. Bi, S. K. Pentyala, K. Ramnath, S. Chaudhuri, S. Mehrotra, Zixu, Zhu, X.-B. Mao, S. Asur, Na, Cheng, A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More, 2024. URL: http://arxiv.org/abs/2407.16216. doi:10.48550/arXiv.2407.16216, arXiv:2407.16216 [cs].

[78] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S. R. Bowman, E. Hubinger, Alignment faking in large language models, 2024. URL: http://arxiv.org/abs/2412.14093. doi:10.48550/arXiv.2412.14093, arXiv:2412.14093 [cs].