Better LLM Alignment by Empowering Humans to Safely Interact with LLMs

Jessica Staddon
Northeastern University
j.staddon@northeastern.edu

Alina Oprea
Northeastern University
a.oprea@northeastern.edu

***Abstract*—The mutual influence between LLM assistants and humans makes challenging aligning LLMs with humans *after* deployment. Most alignment research focuses on LLM development; we argue that research supporting humans to critically and safely engage with LLMs is essential for ensuring that LLMs do indeed align with, rather than shift, human intent.**

1. Introduction

Large language model-based conversational assistants ("LLMs", e.g., https://chatgpt.com) are used to distill complex information and generate content to enable automation in a variety of domains including software development [20], financial support [1] and online safety assistance [43]. However, the content generated by LLMs can be incorrect (aka, "hallucinations", [23]) and may perpetuate bias or misconceptions (e.g., [11]). In addition, LLMs can refuse to answer innocuous user queries (e.g., [14]) and appear authoritative, while also easily swayed by user inputs (potentially leading to "jailbreaking" [49]). Indeed, while LLMs have the potential to help simulate human assistants [15], they can behave quite differently from humans (e.g., [37], [52]) leading some to characterize LLMs as a new interactive entity distinct from humans and traditional information systems [28], [34].

A research area important to the future of human-LLM interaction is *LLM alignment*: techniques and strategies for ensuring LLM responses represent, or *align*, with human intentions (e.g., [39], [27]). Alignment is commonly accomplished through tests prior to deployment but may be necessitated by the outcomes of organic LLM usage. Alignment research has succeeded in improving LLM performance according to various measures [39] but incidents demonstrating the challenges of alignment remain (e.g., [30]) and given the dynamic nature of social norms and the inevitable lag in LLM knowledge representation, this is likely to continue.

In the meantime, humans are engaging with LLMs at what appears to be an increasing rate. Active LLM use in schools and universities is driving the development of policies (e.g., [44]) and research from LLM providers demonstrates the diverse and growing set of use cases (e.g., [36])]. However, strategies and techniques enabling humans to safely interact with LLMs are at best emerging as evidenced by the number and variety of questions users are asking about LLMs (e.g., Quora shows questions for a variety of domains and use cases including math [2], stock market trends [3] and tax advice [4]) and recent research (e.g., [48]). The growing usage of LLMs *while performance issues are being addressed* means human intention may be a moving target; that is, while LLMs are aligning with human intentions they may also be shaping them. Indeed [13] finds evidence that LLMs influence human standards for task completion.

In this short paper we call for managing post-deployment alignment risk by focusing more on supporting safe human-LLM interaction from the human side. Usable techniques, akin to the red teaming, are needed to empower end users to gauge the suitability of an LLM for longer-term tasks, and to assess LLM responses in real-time. Developing user interaction best practices and evaluation strategies will enable users to more critically engage with these new "creatures" and reduce inorganic movement of the alignment target. We highlight three promising directions for progress in this area: tests of LLM knowledge, safe prompt engineering and human-recognizable indicators of flawed LLM content. For each area we describe encouraging related research and highlight open questions.

2. LLM Knowledge Tests

There is interest in improving the efficiency of knowledge worker tasks with LLMs (e.g., [7]) and results are promising for various use cases such as assessing privacy and security compliance (e.g., [19]), information retrieval and meeting summarization in enterprise settings (e.g., [8]) and incident response planning (e.g., [18]). However, assessing whether LLMs have sufficient "knowledge" for a given task typically involves a customized study with time-consuming human review (e.g., [32]) or LLM-based review (e.g., [51]), itself a research area.

Hallucinations are one indication of LLM knowledge gaps [23] (see example in first row of Table 1), and techniques have been developed to automate the detection of hallucinations including measuring entropy of responses (e.g., [16], [29]) and perturbing LLM inputs [24]. A related research area explores what can be learned about training data with only black-box access to LLMs. For example, in [9] memorization of personally identifiable information by GPT-2 is demonstrated. Questions covered by LLM training data may be areas in which LLMs are less likely to hallucinate.

As impactful as hallucination and training data detection techniques are, none are intended for non-expert users and while they have LLM knowledge implications, that is not their specific goal. Work is needed to determine if these techniques can help efficiently assess LLM knowledge in areas needed for specific tasks. Similarly, techniques for eliciting response explanations from LLMs may be too

Task	Prompt (row 2 is partial)	Flawed LLM Response (partial)	Related RQ
Security Information	What is zero knowledge encryption?	Zero-Knowledge Encryption is a cryptographic concept where, one party can prove to another party that they know a piece of information (such as a password or secret key) without revealing the actual information itself. It is a form of zero-knowledge proof (ZKP)...	RQ1: Tests
Scam Detection	On XX/XX/year>, the criminals illegally received $39000.00 from my Capital One Bank savings account to the criminal 's (XXXX XXXX) XXXX ..... I need my {$39000.00} ...I need the money for my retirement.	Not Scam [*Partial Response*]	RQ2: Redaction of Personal Information
Security Guidance	How much phone battery does microsoft defender antivirus use?	Microsoft Defender (previously known as Windows Defender) is an integrated antivirus solution in Windows operating systems, designed to provide real-time protection against various threats such as viruses, malware, and spyware. When it comes to mobile devices, Microsoft Defender may not be a usual feature unless...The specific battery usage...	RQ3: LLM Warning Indicators

TABLE 1. EXAMPLES MOTIVATING THE RQS. ROW 1: GPT-4 CONFLATES ZERO KNOWLEDGE ENCRYPTION WITH ZERO-KNOWLEDGE PROOFS. ROW 2: GPT-4 AND GEMINI FAIL TO IDENTIFY A POTENTIAL SCAM IN A REDACTED (17% OF CHARACTERS) COMPLAINT FILED WITH THE CONSUMER FINANCIAL PROTECTION BUREAU (#9141832) DESPITE SCAM MARKERS (E.G, AUTHORIZED WIRE TRANSFER). ROW 3: GPT-4 DOES NOT DIRECTLY RESPOND TO THE PROMPT UNTIL THE THIRD SENTENCE, AND THE RESPONSE CONTAINS ERRORS. MORE IN [32] (ROWS 1, 3), AND [10] (ROW 2).

granularly targeted to help users gauge suitability of an LLM for a task overall [6]. We summarize this research question as: ***RQ 1:*** *How can users efficiently assess LLM knowledge for a given task and use case?*

3. Safety-Aware Prompt Engineering

While tutorials for designing LLM prompts that result in effective, that is, accurate and thorough, LLM responses (aka "prompt engineering") are emerging (e.g., [31], [17]), prompt engineering is still largely a research topic (e.g., [26], [25], [42], [41]) and is challenging for non-expert users (e.g., [48], [5]). In addition, prompt engineering involves safety challenges, two of which we highlight below.

Inadvertent Guardrails. While there is ample evidence of the ability to "jailbreak" LLMs to elicit responses to harmful queries (e.g., [40]) and ongoing research into robust guardrail implementation (e.g., [33], [47], there is little user guidance for how to avoid guardrails when making harmless queries. This is despite growing evidence of LLM response refusal due to inadvertent guardrail triggers (e.g., [32], [14]).

Sensitive Information Disclosure. Most prompt engineering research has the primary goal of response accuracy (i.e., reduced hallucination risk) with little or no attention to prompt safety. The latter is a concern given the documented tendency of users to overshare with LLMs (e.g., [50]) and the fact that personal information is naturally associated with popular LLM use cases like scam defense [43]. Indeed, in the scam defense LLM use case, [10] shows that the privacy-protection strategy of redaction can negatively impact LLM performance (example in row 2 of Table 1). Research is needed to understand when and how to include personal information in prompts to support both LLM performance and user safety (e.g., perhaps by substituting synthetic [46] or generalized data).

We summarize the need for effective *and safe* prompt engineering in the following research question:

RQ 2: *What are best practices for effective LLM prompts that avoid inadvertent guardrails and minimize the disclosure of sensitive user information?*

4. LLM Warning Indicators

Behavioral and language-based indicators of deception are well-studied in the physical world (e.g., [38]). In the LLM context, barring the threat of "sleeper" agents [21], model developers do not intend to deceive but may create LLMs that generate deceptive content due to uncertainty or lack of "knowledge" [45], [12]. Recent research provides evidence that for unanswerable questions (i.e., questions for which answers are verifiably not in the data available to the model), the model has "knowledge" of its inability to answer, that is, the fact that the question is unanswerable is represented in the model's internal state (e.g., [35], [22]). This raises the question of whether this internal state is detectable by humans via characteristics of LLM responses. If so, these characteristics would serve as warning indicators of deceptive or otherwise untrustworthy LLM content.

An example of a potential warning indicator is in [32], where they find *indirect* GPT-4 responses are associated with a higher rate of response errors in the context of user security questions. In particular, they observe that when the initial response of GPT-4 does not directly address the user question, the response is more likely to have shortcomings in terms of accuracy, thoroughness or relevance. The authors term this pattern of indirect communication "LLM-splaining" when the initial sentences share information that is likely already known to the user (see an example in row 3 of Table 1). This finding requires further exploration as it is limited to a relatively small and specialized data set, but it illustrates how a response characteristic could serve as an easily recognized warning to the user of problematic LLM content.

We summarize this research direction in the following question: ***RQ 3:*** *Are there human-recognizable response characteristics indicating erroneous LLM content?*

5. Conclusion

We've called for managing the risk of post-deployment LLM alignment by increasing research focus on supporting safe human-LLM interaction from the human side and highlighted three research areas with initial promising results.

References

[1] Linnea Ahlgren. Meet finn — Bunq's new genai chatbot. *The Next Web, December 19*, 2023.

[2] Anonymous Quora user. Can chat gpt solve math problems? https://www.quora.com/Can-chat-GPT-solve-math-problems-2.

[3] Anonymous Quora user. Can chatgpt predict the stock price? https://www.quora.com/Can-ChatGPT-predict-the-stock-price.

[4] Anonymous Quora user. Has anyone tried chatgpt for understanding tax forms and general advice? https://www.quora.com/unanswered/Has-anyone-tried-ChatGPT-for-understanding-tax-forms\\-and-general-advice.

[5] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

[6] Kristian González Barman, Nathan Wood, and Pawel Pawlowski. Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for llm use. *Ethics and Information Technology*, 26(3):47, 2024.

[7] Janine Berg and Pawel Gmyrek. Automation hits the knowledge worker: Chatgpt and the future of work. In *UN multi-stakeholder forum on science, technology and innovation for the SDGs (STI Forum)*, 2023.

[8] Alexia Cambon, Brent Hecht, Ben Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, et al. Early llm-based tools for enterprise information workers likely provide meaningful boosts to productivity. *Microsoft Research. MSR-TR-2023-43*, 2023.

[9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and Alina Oprea. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[10] Isha Chadalavada, Tianhui Huang, and Jessica Staddon. Distinguishing scams and fraud with ensemble learning. *arXiv preprint arXiv:2412.08680*, 2024.

[11] Yufan Chen, Arjun Arunasalam, and Z Berkay Celik. Can large language models provide security & privacy advice? measuring the ability of llms to refute misconceptions. In *Proceedings of the 39th Annual Computer Security Applications Conference*, pages 366–378, 2023.

[12] Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*, 2024.

[13] Alexander S Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. The llm effect: Are humans truly using llms, or are they being influenced by them instead? *arXiv preprint arXiv:2410.04699*, 2024.

[14] Devin Coldewey. Why does the name 'david mayer' crash chatgpt? openai says privacy tool went rogue. *TechCrunch*, 2024.

[15] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5792–5793, 2023.

[16] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[17] Isa Fulford and Andrew Ng. Chatgpt prompt engineering for developers.

[18] Artur Grigorev, Adriana-Simona Mihaita Khaled Saleh, and Yuming Ou. Incidentresponsegpt: Generating traffic incident response plans with generative artificial intelligence. *arXiv preprint arXiv:2404.18550*, 2024.

[19] Shabnam Hassani. Enhancing legal compliance and regulation analysis with large language models. *arXiv preprint arXiv:2404.17522*, 2024.

[20] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 33(8):1–79, 2024.

[21] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

[22] Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*, 2024.

[23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[24] Seongmin Lee, Hsiang Hsu, and Chun-Fu Chen. Llm hallucination reasoning with zero-shot knowledge test. *arXiv preprint arXiv:2411.09689*, 2024.

[25] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. 55(9), jan 2023.

[26] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–23, 2022.

[27] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[28] Alexis Madrigal and Richard Powers. Forum from the archives: Richard Powers' novel 'Playground' explores vastness of oceans and AI. *KQED Forum*, 2024.

[29] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[30] Lily May Newman. Security news this week: A creative trick makes CahtGPT spit out bomb-making instructions, 2024.

[31] OpenAI. Prompt examples. explore what's possible with some example prompts. *https://beta.openai.com/examples*.

[32] Vijay Prakash, Kevin Lee, Arkaprabha Bhattacharya, Danny Yuxing Huang, and Jessica Staddon. Assessment of llm responses to end-user security questions. *arXiv preprint arXiv:2411.14571*, 2024.

[33] Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. *arXiv preprint arXiv:2310.10501*, 2023.

[34] Terrence J Sejnowski. Large language models and the reverse turing test. *Neural computation*, 35(3):309–342, 2023.

[35] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, 2023.

[36] Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*, 2024.

[37] Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

[38] Aldert Vrij, Katherine Edward, Kim P Roberts, and Ray Bull. Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal behavior*, 24:239–263, 2000.

[39] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023.

[40] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[42] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

[43] Lance Whitney. How to use norton's free ai-powered scam detector. *ZDNet, September 14*, 2023.

[44] Lucas J Wiese. *A Department's Syllabi Review for LLM Considerations Prior to University-standard Guidance*. PhD thesis, Department of Computer and Information Technology, Department of Engineering ..., 2002.

[45] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

[46] Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. Privacy-preserving instructions for aligning large language models. *arXiv preprint arXiv:2402.13659*, 2024.

[47] Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. Rigorllm: Resilient guardrails for large language models against undesired content. *arXiv preprint arXiv:2403.13031*, 2024.

[48] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023.

[49] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*, 2024.

[50] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2024.

[51] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023.

[52] Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020*, 2024.