ReaLLM

Interface-Level Transparency as a Safety Intervention in Human–LLM Interaction

Mohsen HassanNejad
AISES Course Project — Track 1B: Technical Research with Write-Up

Presentation Deck | Demo | Paper | GitHub

1. Introduction

The AI safety field has invested heavily in technical interpretability. AI safety research is focused on the transparency and explainability of AI models. Yet, explainable AI (XAI) approaches have been predominantly algorithm-centered, focused on revealing model internals rather than the social context in which AI operates (Ehsan et al., 2021). Mechanistic interpretability reverse-engineers neural circuits; representation engineering reads high-level model representations (Center for AI Safety, 2024). These are important but researcher-facing. Millions of everyday users do not benefit from knowing about polysemantic neurons.

This paper argues that for large language models (LLMs), more accessible layers of social transparency have been neglected, ones written in plain language. System prompts are one such mechanism used by different AI labs. Essentially, guidelines that define model identity, encode prohibitions, and establish guardrails to influence their behavior and responses. However, there is a caveat: these instructions, unbeknownst to the user, are attached to their request before it reaches the AI model. Research has documented a "gulf between user expectation and experience" with conversational agents for nearly a decade (Luger & Sellen, 2016); concealed system prompts fit in part of that frame.

In a prior article, I analyzed system prompts from major AI labs. This paper builds on that analysis: interface-level transparency is not a replacement for algorithm-centered interpretability, but rather a complementary mechanism that addresses the socio-organizational context shaping human-AI interaction.

2. What LLM Interfaces Hide

Analysis of system prompts from major AI labs reveals three categories of hidden control:

Identity and Persona. Models are assigned specific personalities. Some adopt a "social chameleon" approach, mirroring user tone; others are configured as "careful and precise" or "deep research assistants."

Behavioral Constraints. System prompts contain layered prohibitions: hard guardrails against weapons information; intellectual guardrails limiting quotations; accuracy directives against hallucination. These operate silently, shaping what users can elicit.

Meta-Guardrails. Most tellingly, models are instructed to conceal their own instructions. OpenAI's prompts forbid disclosure of system messages; Claude is told to withhold reasoning behind refusals "to avoid irritating the user." The system hides its own hiding.

Studies such as Greshake et al. (2023) show that LLM applications "blur the line between data and instructions". Not knowing that a complicated constitution of instructions gets attached to your prompts when you interact with an LLM has direct and indirect consequences. For instance, users hold simplified, often incorrect mental models of LLM ecosystems (Wang et al., 2025).

Indirectly, the lack of understanding exposes users to other risks catalogued by the MIT AI Risk Repository (Slattery et al., 2024): Domain 5.1 identifies "overreliance and unsafe use", users trusting AI without understanding its constructed nature. Domain 7.4 flags opacity as creating mistrust and an inability to identify and correct errors.

The irony is that system prompts are written in natural language and are, in a sense, the most accessible layer of model architecture. On the other hand, it is not so clear as to how these system-level instructions should be communicated to the users and if they could even be all that useful. The next section addresses these tensions.

3. Interface Transparency as a Soft Safety Mechanism

This paper advances a design thesis: selectively exposing hidden layers, the model’s given roles, constraint rationales, and uncertainty signals could function as a soft safety intervention, recalibrating user mental models without modifying the underlying model.

This is inference level explainabililty. Technical interpretability asks: *What is the model doing internally?* Interface transparency asks: *What does the user need to know?* The second directly serves the millions making decisions based on AI outputs.

Transparency is not automatically beneficial. Poursabzi-Sangdeh et al. (2021) found interpretable "clear box" models led to *increased* overreliance—transparency created an illusion of understanding. Buçinca et al. (2021) showed explanations did not reduce overreliance; users developed general heuristics rather than evaluating each recommendation.

Yet transparency can work. Vasconcelos et al. (2023) found explanations reduce overreliance when tasks are difficult and disclosure reduces verification costs. The implication is that design matters. Selective, context-sensitive disclosure outperforms both maximal transparency and total opacity.

Trade-offs remain: cognitive load from excessive disclosure; performative transparency creating false confidence; users gaming revealed prompts; the question of who decides what gets revealed. These argue for intentional design, not abandoning transparency.

4. Proof-of-Concept Prototype

To bring these claims and tensions to attention, I developed a minimal prototype. It is a conceptual demonstrator, not an evaluated system.

The prototype implements two features:

1. Full prompt disclosure — Users can view the complete system prompt.
2. Transparency panel — A side panel shows which parts of the system prompt may be influencing the current response.

The prototype illustrates how identical outputs can be interpreted differently when contextual layers become visible. A response reading as authoritative in a standard interface becomes provisional when accompanied by uncertainty markers.

5. Conclusion

Contribution. This paper argues for socially-situated transparency, with system prompts as a prime candidate. Ehsan et al.'s (2021) 4W framework asks: Who made decisions about the system? What does it do? When? Why? Disclosing system prompts address these questions directly. The infrastructure for this work exists: Amershi et al. (2019) synthesized 18 validated guidelines for human-AI interaction, including capability disclosure and confidence communication. The opportunity and challenge remain in applying them to LLMs.

Limitations. Many users may not want this transparency; for casual use, it could feel like friction. Anthropic's system prompt exceeds 10,000 words; summarizing it without overwhelming users is a design problem in itself. Even plain-language prompts contain concepts (guardrails, persona engineering) that require context to understand. How do users verify a disclosed prompt is complete? Partial disclosure could be worse than none. Our prototype's explainer model also requires a second LLM call, which creates computational overhead that is difficult to justify at scale.

Future work. Empirical questions remain: how does system prompt visibility affect user reliance? Can disclosure be adaptive to context and expertise? How can they best be communicated? What policy standards should govern transparency? All of these questions require attention and further analysis.

Concealing system prompts is a design choice, not a technical necessity, and that choice has safety implications as well as long-term user impact. Interface-level transparency will not solve alignment, but it offers a tractable, user-facing complement to the more technically complex work the field is wrestling with.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Article 3, 1–13. https://doi.org/10.1145/3290605.3300233

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188, 1–21. https://doi.org/10.1145/3449287

Center for AI Safety. (2024). *Introduction to AI safety, ethics, and society*. https://www.aisafetybook.com/

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 82, 1–19. https://doi.org/10.1145/3411764.3445188

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90. https://doi.org/10.1145/3605764.3623985

Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 237, 1–52. https://doi.org/10.1145/3411764.3445315

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). *The AI Risk Repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence*. https://doi.org/10.48550/arXiv.2408.12622

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), Article 129, 1–38. https://doi.org/10.1145/3579605

Wang, X., Wang, X., Park, S., & Yao, Y. (2025). Mental models of generative AI chatbot ecosystems. *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1016–1031. https://doi.org/10.1145/3708359.3712125