Responsible Prompting Recommendation in Multi-Turn Interaction with LLMs

Vagner Figueredo de Santana[1,3], Ashwath Vaithinathan Aravindan[1], Cassia Sampaio Sanctos[2], Tiago Machado[2], Luan Soares de Souza[2]

[1]IBM Research - Yorktown Heights, NY, USA

[2]IBM Research - São Paulo, SP, Brazil

[3]NIC.br - São Paulo, Brazil

{vsantana,csamp,tiago.machado,luanssouza}@ibm.com, vaashwath@gmail.com

***Abstract. Introduction**: Large Language Models (LLMs) are being proposed as a solution to be applied in multiple workflows, but they lack proper user guidance and Responsible AI awareness in prompting-time, i.e., before sending a given prompt to an LLM. **Objective:** In this context, this research proposes a way to provide user guidance and Responsible AI awareness while people interact with LLMs in a multi-turn fashion. **Methodology or Steps**: In our tool, users receive Responsible AI suggestions while writing prompts in multi-turn conversations with LLMs. **Results:** We expect this work motivates more prompting recommender systems aiming at speeding up prompting tasks, users' agency, transparency, and promoting Responsible AI in prompting-time.*
***Keywords** Responsible AI, Artificial Intelligence, Social Values.*

***Resumo. Introdução:** Large Language Models (LLMs) estão sendo propostos como solução para múltiplos fluxos de trabalho, mas eles comumente não contam com orientação adequada a usuários e nem fornecem informações sobre IA Responsável no momento da criação de prompts. **Objetivo:** Nesse contexto, esta pesquisa propõe uma forma de orientar e conscientizar pessoas sobre IA Responsável enquanto interagem com LLMs em múltiplos turnos. **Metodologia ou Etapas**: Em nossa ferramenta, usuários recebem sugestões de IA Responsável enquanto escrevem prompts para LLMs. **Resultados:** Espera-se que este trabalho motive mais sistemas de recomendação de prompting para promover IA Responsável no momento da criação dos prompts.*
***Palavras-Chave** IA Responsável, Inteligência Artificial, Valores Sociais.*

1. Introduction

Over the last decade, Responsible Innovation initiatives have highlighted the importance and necessity of proactively and systematically considering harms and benefits across multiple technologies. However, the importance of responsible artificial intelligence (AI) *specifically* has emerged as a 'must have' due to recent advances in Generative AI (GenAI) and associated Large Language Models (LLMs). In this context, 'Responsible AI' (RAI) can likewise be seen as an umbrella term for initiatives that work to ensure appropriate business and societal choices when adopting, building, and deploying AI, encompassing research, responsibilities, and practices that create positive, accountable, and ethical AI development and operation [Perri 2023].

Because of GenAI's stochasticity and variability [Weisz et al. 2023] and the multiple and inherent difficulties of prompting well (e.g., efficiently and sufficiently) [Zamfirescu-Pereira et al. 2023], Prompt Engineering has emerged as a new and dedicated activity, role, and interaction modality. Prompt Engineering (aka *prompting*) can be defined as *"the process of* ***communicating effectively*** *with an AI to achieve* ***desired results****"* [Learning Prompting 2023]. Since GenAI may lead to a variety of well-documented harms - including but not limited to erasing or obfuscating social terms or issues, stereotyping or misrepresenting people, and/or negatively impacting people's agency and well-being [Blodgett et al. 2022] - there is a need to combine existing RAI towards prompting as a specific and vital practice in this space. In this context, Responsible Prompting was defined as *the process of communicating effectively with an AI system to achieve desired results while avoiding or minimizing harms, promoting responsible practices, and employing mechanisms for anticipation, reflexivity, inclusion, and responsiveness* [Santana et al. 2025]. Hence, this work contributes to the field of Human-Computer Interaction (HCI) by showing how responsible prompting can be employed in multi-turn interactions, in a lightweight and LLM-agnostic way.

2. Related Work

Prompting is a relatively new way of interacting with AI, considered by some professionals to even be an 'artform' [Bhatti 2023, Chang et al. 2023, Beauchemin 2023] due to the ways in which users must creatively navigate GenAI's inconsistent and imperfect outputs [Weisz et al. 2023]. As with any emerging technology being quickly adopted at a global scale, it is quite difficult to properly measure and track GenAI's societal impacts. Prompts and their results (model outcomes) are being sold as data assets in and of themselves in various marketplaces (e.g., Promptbase[1], Etsy[2]), prompt templates are being shared openly, freely, and at-scale in certain communities [Chang et al. 2023], and datasets of prompts such as Awesome ChatGPT Prompts[3] and AttaQ [Kour et al. 2023] have been open-sourced for people to test and assess various LLMs. However, there are currently no standards for assessing the quality of these prompts or many of their outcomes [Maia Polo et al. 2024], and the plurality of prompting resources neither necessitates nor guarantees that users will learn how to intentionally prompt GenAI responsibly.

Regarding prompting specifically, online references [Learning Prompting 2023, Melanson e Maman 2023, Simonovsky 2023, Sarikas 2023] and books [Gallery 2022, Diab et al. 2022] provide initial recommendations about how to better obtain desired model results, such as the **3 principles format** (be specific, provide context, and iterate & improve), the **RGC Style** (Role, Result, Goal, Context, Constraint), the **CLEAR** framework (conciseness, logic, explicitness, adaptability, and reflectiveness) [Lo 2023], or the **"anatomy of prompts"** [Santana 2024]. Practices, guardrails, and defensive tactics are also actively being developed to identify and prevent for adversarial prompting attacks [Balas et al. 2024].

In this vein, multiple tools are also being proposed or created to guide users through prompting considerations. Existing approaches include:

[1] https://promptbase.com/
[2] https://www.etsy.com/
[3] https://github.com/f/awesome-chatgpt-prompts

integrated development environments (IDEs) for prompting [Fiannaca et al. 2023], prompt editing tools [Wang et al. 2023], tools for supporting test-driven prompt engineering [Beauchemin 2023], tools that leverage LLMs to generate synthetic prompts [Zhou et al. 2022], tools for helping users on prompt template chaining [Arawjo et al. 2024, Anthropic 2024], tools to support programmers to work collaboratively when generating prompts for coding assistance [Feng et al. 2024, Cohere 2024], tools for supporting users in communicating intentions to a text-to-image model [Brade et al. 2023], and systems and methods for visually exploring prompting elements based on generated content, including domain knowledge terms [TensorFlow 2023, Promptomania 2023, Saxifrage 2023, Character.AI 2024, PicFinder 2023], knowledge graph [Jiang et al. 2023], and associated embedding spaces [Rost e Andreasson 2023, Brade et al. 2023]. While the tools listed here have different degrees of prompt automation, there are gaps present, particularly when considering the power and potential of prompting as both an interface and action to enhance RAI through awareness and daily practice.

3. Multi-Turn Responsible Prompting Recommendation

The responsible prompting recommender system was designed to be an LLM-agnostic tool used in prompting-time, i.e., before the prompt is actually sent to the GenAI, while the user is writing the prompt. Any lightweight sentence transformer providing an endpoint for sentence embeddings can be used for this solution (e.g., all-minilm-l6-v2). Moreover, the prompt then can be sent to any LLM, for instance, *mistral-7b-instruct*, *llama-4-scout*, etc. (Figure 1).

Figura 1. Value-based prompt recommendation for a given input prompt. Values recommended (light green) include measurability, money, progress, reliability, and appropriate. The latter was selected and respective sentence was added to the prompt (dark green).

The recommendations are based on a dataset of human-curated sentences residing in a JSON file. The current dataset of human-curated sentences consists of +2500 sentences, including 'positive' sentences aiming at adding beneficial social values to a user's prompts, as well as 'negative' adversarial sentences used as references to prevent harmful prompts to be sent to the model. The JSON file was structured as follows: (1) into two blocks of sentences (positive and negative) to prevent sentences with similar semantics but opposing valence to be clustered together; (2) into clusters of sentences based on the associated positive/negative values. Clusters were created to allow the similarity search to be performed in two steps: first through the clusters' centroids, and then for the most similar sentence in the cluster. The tool resulting from this research is being open-sourced[4] so the HCI community can benefit and contribute to project's code

[4]https://github.com/IBM/responsible-prompting-api

base, sentences, and values, making room for more pluralistic social values and up-to-date adversarial sentences.

The goal of the proposed recommender system is to suggest sentences to be added or removed from an input prompt, acting as a user guidance for how to embed social values within prompts while preventing known harms. From any given input text, the algorithm splits the prompt into sentences and employ a similarity search for finding relevant prompt sentences to recommend and input sentences to remove (Figure 2). Finally, by employing the responsible prompting recommendation in multi-turn conversations, we aim at embedding Responsible AI while people interact with LLMs, before content is actually generated (Figure 3).

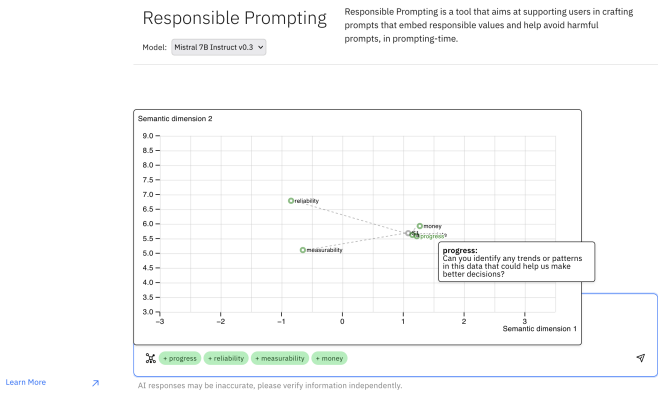

Figura 2. Explainability feature showing the role that embeddings' similarities play in the recommendations. Each node represents embeddings for sentences entered (S1, S2, etc.) and recommendations (progress, reliability, measurability, etc.).

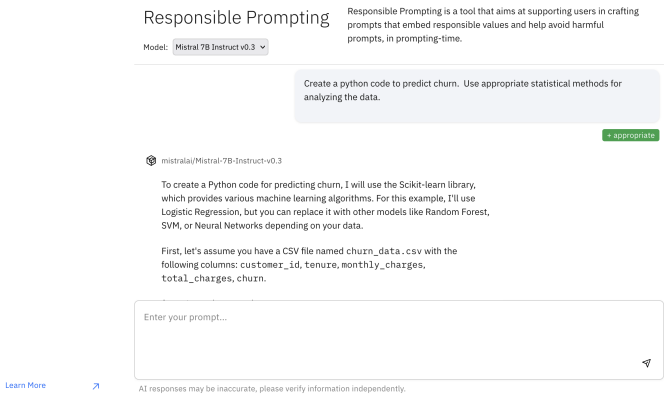

Figura 3. Outcome provided by the selected LLM (mistral-7b-instruct) taking into consideration the input prompt and the selected recommendation.

4. Discussion

This research contributes with a lightweight LLM-agnostic approach to provide recommendations of good practices and prevention of harmful ones, in prompting time. The proposed recommender system differs from the tools detailed in the related work

section in the following key aspects: it does not require any effort from the user, given that the templates for responsible prompting are chained automatically based on the users' input (in real-time); given that the recommendation is automatic, our UI is more focused on visual cues for highlighting sentences to be added or removed; the system provides automatic recommendations to increase prompt orientation towards critical RAI considerations among professionals; it not only suggests additions to prompts but also recommends the removal of sentences that may trigger harmful responses from the model; our system follows a lightweight approach instead of training a model from scratch; and our tool does not change the users' input without their consent. This practice has multiple benefits. It not only supports more responsible prompts, but enables the user to reflect on the prompts being written in a dynamic way, making critical thinking part of human-AI interaction. By improving the process of prompting, there is a downstream effect in which the model might become more responsible, once it is retrained with the input information, i.e., the data in the pretrained model will eventually encompass more responsible values, which will also make the model's output more responsible through time. Therefore, benefiting the development of responsible models and the community around it.

5. Ethical Concerns

For this demo, we did not conduct studies involving human participants. This project is open-sourced and more information about contributions, contributors, license, and code of conduct can be found on project's repository[5].

Referências

Anthropic (2024). Promptgenerator. `https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator`. Accessed: 2025-08-21.

Arawjo, I., Swoopes, C., Vaithilingam, P., Wattenberg, M., e Glassman, E. L. (2024). Chainforge: A visual toolkit for prompt engineering and llm hypothesis testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Balas, M., Wong, D., e Arshinoff, S. (2024). Artificial intelligence, adversarial attacks, and ocular warfare. *AJO International*, 1(3).

Beauchemin, M. (2023). Mastering ai-powered product development: Introducing promptimize for test-driven prompt engineering. `https://shorturl.at/cVH08`. Accessed: 2025-08-21.

Bhatti, B. M. (2023). The art and science of crafting effective prompts for llms. `https://shorturl.at/vsUUK`. Accessed: 2025-08-21.

Blodgett, S. L., Liao, Q. V., Olteanu, A., Mihalcea, R., Muller, M., Scheuerman, M. K., Tan, C., e Yang, Q. (2022). Responsible language technologies: Foreseeing and mitigating harms. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

[5]`https://github.com/IBM/responsible-prompting-api`

Brade, S., Wang, B., Sousa, M., Oore, S., e Grossman, T. (2023). Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Chang, M., Druga, S., Fiannaca, A. J., Vergani, P., Kulkarni, C., Cai, C. J., e Terry, M. (2023). The prompt artists. In *Proceedings of the 15th Conference on Creativity and Cognition*, C&C '23, page 75–87, New York, NY, USA. Association for Computing Machinery.

Character.AI (2024). Promptpoet. `https://github.com/character-ai/prompt-poet`. Accessed: 2025-08-21.

Cohere (2024). Prompttuner. `https://docs.cohere.com/docs/prompt-tuner`. Accessed: 2025-08-21.

Diab, M., Herrera, J., Sleep, M., Chernow, B., e Mao, C. (2022). *Stable Diffusion Prompt Book*. OpenArt.

Feng, L., Yen, R., You, Y., Fan, M., Zhao, J., e Lu, Z. (2024). Coprompt: Supporting prompt sharing and referring in collaborative natural language programming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Fiannaca, A. J., Kulkarni, C., Cai, C. J., e Terry, M. (2023). Programming without a programming language: Challenges and opportunities for designing developer tools for prompt programming. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Gallery, D. (2022). *DALL-E2 Prompt Book*. Dallery Gallery.

Jiang, P., Rayan, J., Dow, S. P., e Xia, H. (2023). Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Kour, G., Zalmanovici, M., Zwerdling, N., Goldbraich, E., Fandina, O. N., Anaby-Tavor, A., Raz, O., e Farchi, E. (2023). Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*.

Learning Prompting (2023). Prompt engineering guide. `https://learnprompting.org/docs/intro`. Accessed: 2025-08-21.

Lo, L. (2023). The art and science of prompt engineering: A new literacy in the infomration age. *Internet Reference Services Quarterly*, (4):203–210.

Maia Polo, F., Xu, R., Weber, L., Silva, M., Bhardwaj, O., Choshen, L., de Oliveira, A., Sun, Y., e Yurochkin, M. (2024). Efficient multi-prompt evaluation of llms. *Advances in Neural Information Processing Systems*, 37:22483–22512.

Melanson, J. e Maman, B. (2023). Chatgpt +25 powerful ai tools 10x your productivity & creativity | chatgpt, generative ai, prompt engineering, dall-e2. E-learning Course.

Perri, L. (2023). What's new in artificial intelligence from the 2023 gartner hype cycle. `https://shorturl.at/xCZee`. Accessed: 2025-08-21.

PicFinder (2023). Picfinder. https://picfinder.ai/. Accessed: 2025-08-21.

Promptomania (2023). Generic prompt builder. https://promptomania.com/generic-prompt-builder/. Accessed: 2025-08-21.

Rost, M. e Andreasson, S. (2023). Stable walk: An interactive environment for exploring stable diffusion outputs. *Proceedings of the 4th Workshop on Human-AI Co-Creation with Generative Models - HAI-GEN 23*.

Santana, V. F. D. (2024). Challenges and opportunities for responsible prompting. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Santana, V. F. d., Berger, S. E., Candello, H., Machado, T., Sanctos, C. S., Su, T., e Williams, L. (2025). Responsible prompting recommendation: Fostering responsible ai practices in prompting-time. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Sarikas, O. (2023). Midjourney ai - best prompt ticks - beginners guide - beginners - mj explained - nft art. https://www.youtube.com/watch?v=lFI8JQvPfu8. Accessed: 2025-08-21.

Saxifrage (2023). Visual prompt builder. https://tools.saxifrage.xyz/prompt. Accessed: 2025-08-21.

Simonovsky, T. (2023). Chatgpt for data science and data analysis in python. https://www.udemy.com/course/chatgpt-for-data-science-and-data-analysis-in-python/. Accessed: 2025-08-21.

TensorFlow (2023). Embedding projector. https://projector.tensorflow.org/. Accessed: 2025-08-21.

Wang, Y., Shen, S., e Lim, B. Y. (2023). Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–29.

Weisz, J. D., Muller, M., He, J., e Houde, S. (2023). Toward general design principles for generative ai applications. https://arxiv.org/abs/2301.05578. Accessed: 2025-08-21.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., e Yang, Q. (2023). 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*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., e Ba, J. (2022). Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*.