Exploring User Prompting Behavior in LLM Interactions

Maximilian Slapnik
Maximilian.Slapnik@campus.lmu.de
LMU Munich
Munich, Germany

ABSTRACT

Artificial Intelligence (AI) plays an increasingly important role in the daily lives of millions of people. Large Language Models (LLMs) are the most prominent implementation of AI that is used not only by experts, but equally by ordinary users as well. LLMs can respond to any textual input (prompts) with human-like answers, leveraging the training data that was used to implement the model. Even though prompting LLMs seems very straightforward, the question arises if it is possible to streamline the interactions with said models in order to optimize outputs. We explore the behavior of a randomized trial of 100 interactions of users with LLMs that are publicly available on ShareGPT. The goal of this investigation is the discovery of recurring patterns in behavior and the evaluation of human tendencies as well as biases of users when interacting with AI models in order to understand current behaviors and propose optimization opportunities.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design**; • **Information systems** → *Information retrieval*; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

Large Language Models, user behavior, prompting, interaction patterns

1 INTRODUCTION

2 BACKGROUND AND RELATED WORK

2.1 Large Language Models (LLMs)

∘ General information on LLMs, such as their workings, training data, text generation, real world usage, and current limitations

2.2 User Interaction with LLMs

∘ Explanation of Prompting
∘ Description of LLM use cases and related work, primarily paying attention to ordinary frequent users (and not only experts)

3 STUDY ON USAGE PATTERNS OF LLM USERS

3.1 Intro and Research Objective

∘ Overview of the study goal, the methodology, and the individual steps that will be taken

3.2 Research Method: ShareGPT

∘ Information on the ShareGPT platform, its user base, its suitability for the study, and which data we are going to use

3.3 Study Results

3.3.1 Findings. ∘ Listing of the results of the study, potentially segregated into categories that can be defined in advance

3.3.2 Observable Trends. ∘ Objective analysis of results with a particular focus on observable trends in user behavior and data patterns (including visualizations such as charts)

4 DISCUSSION

4.1 Observed Behaviour (Synthesis)

∘ Subjective evaluation of findings

4.1.1 Why do users interact with LLMs the way they do? ∘ Reasoning and informed assumptions on the causes of observed behavior

4.1.2 Prompt Improvement Possibilities. ∘ Proposition of ways to enhance prompts as well as associated results based on findings from related research

4.2 Outlook and Future Developments

4.2.1 Auto-GPT. ∘ Introduction to future developments in the realm of LLM interaction, such as AI-based agents which may execute prompts autonomously in the future

4.2.2 Prompt Engineering. ∘ Focus on the newly emerging discipline of prompt engineering which is a direct result of the increased significance of LLMs and required competencies for successful interaction

5 CONCLUSION

6 INTRODUCTION

Artificial Intelligence (AI) -based tools continually gain prominence as regularly leveraged tools in the daily lives of millions of people. In addition to typical AI applications such as recommendation systems or autonomous agents, generative models are notably increasing in popularity as well. One of the most widely used implementations of generative models are Large Language Models (LLMs), the most popular example at the moment being OpenAI's ChatGPT [8]. These models mainly come in the form of text generating chatbots that can answer seemingly any question a user might pose. Nevertheless, it is challenging to optimize the output of the model, since it can vary depending on the user input. Any form of such input to an LLM, whether it is in the form of a task or a question, is commonly referred to as "prompting" the model. Due to the vast application possibilities and promising future developments of LLMs, an exploration of user prompting behavior in interactions with these models is of particular interest.

In this paper, we are going to explain the fundamentals and workings of LLMs and prompting, describe related research in the realm of user - LLM exchange, and perform our own investigation of user behavior in these interactions. This investigation will provide an improved understanding of existing challenges users face when dealing with such models, as well as highlight optimization potential in order to enhance generated output.

Plenty of research has been conducted in the field of user interactions with LLMs.

Since the main part of this work will contain an analysis of real world examples, the reader can expect to gain a better understanding of actual user prompting behavior. To obtain these insights, we will leverage input data mainly gathered from the website ShareGPT, which enables users to store conversations they have had with the ChatGPT model and share them with others.

In the concluding section of this paper we will summarize our findings and explain how and in which way we can recognize findings from related research in our own data samples.

7 BACKGROUND AND RELATED WORK

This section provides an overview over fundamental concepts that are necessary to facilitate comprehensibility of all latter parts. We start with explaining Large Language Models in general, followed by more specific explanations of distinct prompting approaches. These explanations are followed by a more general overview of related work on the subject.

7.1 Large Language Models (LLMs)

One of the most widely used applications areas of generative AI are Large Language Models (LLMs). Among LLMs, the most widely adopted is ChatGPT [8], which is a conversational model being developed by OpenAI. The model is currently publicly accessible and free of charge.

LLMs can be leveraged for a variety of tasks, but their main focus area is Natural Language Processing (NLP). Therefore, most LLMs designed for end users are implemented in the form of chatbots, as is the case with ChatGPT for example. They typically consist of an interface comprised of an input field for the user to type in arbitrary text, as well as an output section that displays generated responses of the model.

Large Language Models are a recent advancement that followed the development of the original transformer architecture, which is a deep learning approach first introduced by researchers in 2017 [14]. In future iterations, the Generative Pre-Training (GPT) [9] approach was adapted for text-based models in particular, laying the foundation for today's most popular conversational LLMs, such as ChatGPT. Since our research revolves around user interaction with dialog-focused models, we will not go into more detail about other application and development areas of LLMs from hereon. In addition to ChatGPT, there are also a variety of similar other models focused on text generation, such as LaMDA [13], Sparrow [4], or BlenderBot 3 [12].

Large Language Models are a central part of our research as we investigate user behavior in LLM conversations. Currently, there are no binding guidelines on the concrete structure when prompting such models, and users are therefore completely unconfined in their way of interacting with them.

7.1.1 Zero-Shot Learning. It is important to differentiate possible forms of prompting that exist in LLM interactions in order to be able to correctly analyze and assess prompts in later parts of this paper. To do so, we will clarify the terms "Zero-Shot-" as well as "Few-Shot Learning". Zero-Shot Learning has been described as "[...] learning how to recognise new concepts by just having a description of them" [10, p. 1].

In regard to prompting, Zero-Shot refers to any form of prompt where the user does not specify a particular example as part of the query [3, p. 1]. An example for such a query could be "Give me five examples of papers that deal with the topic of LLMs". As expected, model performance (accuracy) was measured as inferior in zero-shot prompting settings compared to other settings where more information is provided in the prompt, such as a Few-Shot approach [1, p. 5].

7.1.2 Few-Shot Learning. In Few-Shot Learning in contrast, examples are given, albeit not many. It is generally aimed at ensuring good learning performance with only a few (most often less than 20) supervised examples [11, p. 1]. In a Few-Shot prompting setting, the user accordingly does provide increased guidance for the model. Doing so can help improve model outputs [3, p. 1]. Solely giving one demonstration in the prompt is commonly referred to as a "One-Shot" approach [1, p. 6], whereas multiple demonstrations make it "Few-Shot". To exemplify the differences in learning capabilities depending on the prompting setting, we present an example inspired by the approach of Brown et al. [1] and recent online resources [2] in Table 1. The table displays an interaction we have carried out with the current ChatGPT-3 model, where we instructed the model to classify the sum of a sequence of numbers as either even or odd.

Table 1: Odd Number Sum Evaluations

Prompt:	The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.
A:	*The odd numbers in this group add up to an even number: 15, 5, 13, 7, and 1. Their sum is 41, which is indeed an even number.*
Prompt:	The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1. The answer is False. The odd numbers in this group add up to an even number: 17, 10, 19, 4, 8, 12, 24. The answer is True. The odd numbers in this group add up to an even number: 16, 11, 14, 4, 8, 13, 24. The answer is True. The odd numbers in this group add up to an even number: 17, 9, 10, 12, 13, 4, 2. The answer is False. The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.
A:	*The answer is False. The odd numbers in this group (15, 5, 13, 7, and 1) add up to 41, which is an odd number, not an even number.*

Using this example prompt, we could recreate the effects observed by other researchers. We can witness that the request of the user is indeed first wrongly answered by the model using a zero-shot approach. Only when relying on few-shot prompting and supplying ChatGPT-3 with additional information, the model generates the correct result.

We should note however, that ChatGPT is a dedicated language model, and we have used ChatGPT-3. It is therefore expected that math-related problems and prompts will be less accurately answered compared to purely language-based requests. Depending on the model training it is imaginable that future iterations, such as ChatGPT-4, will presumably improve performance and may therefore not have to rely on few-shot assistance as much.

7.2 User Interaction with LLMs

The majority of existing related work focuses on strategies for designing prompts, challenges users may face in the course of doing so, as well as investigations of user behavior when performing search queries in general. These topics are of particular relevance for our work since we are not only interested in ways to improve prompt outputs, but end user behavior when doing so, too.

7.2.1 Effectiveness of Query Reformulation. Interest in query optimization in order to improve output is widely prevalent ever since the broad availability of large search engines, such as Google [5] or Bing [7]. Huang and Efthimiadis [6] investigated user behavior when reformulating search queries, i.e. modifying a previous query in order to improve results, in as early as 2009. Even though their focus was on search queries alone, it was significant that there were improvements that could be seen after reformulating queries, concluding that "Most reformulation strategies result in some benefit to the user" [6, p. 1].

7.2.2 Few-Shot Learning Capabilities. As indicated in the previous section, the learning setting of an LLM can significantly influence its performance. Related research by Brown et al. [1] for example has explored differences in model performance depending on examples provided as part of the LLM prompt [1]. To do so, they have conducted various experiments directly comparing zero-, one-, and few-shot learning. In conclusion, the trials revealed that LLMs indeed show remarkable performance in various areas, ranging from translation and question answering to reasoning tasks. Notably, few-shot learning models are able to adapt to new tasks with few training data. It becomes clear that it is important to provide adequate, fitting examples as part of a prompt in order to enhance model outcomes.

7.2.3 Challenges in Non-expert Prompt Design.

7.2.4 Opportunities and Challenges for Interactive Prompt Design Applications.

REFERENCES

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165 arXiv:2005.14165 [cs].

[2] DAIR.AI. 2023. Few-Shot Prompting | Prompt Engineering Guide. https://www.promptingguide.ai/techniques/fewshot

[3] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. http://arxiv.org/abs/2209.01390 arXiv:2209.01390 [cs].

[4] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. http://arxiv.org/abs/2209.14375 arXiv:2209.14375 [cs].

[5] Google. 2023. Google. https://www.google.com/?client=safari&output=search&gbv=1&sei=9FJ4ZISCA7-Rxc8P162ouAM

[6] Jeff Huang and Efthimis N. Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, Hong Kong China, 77–86. https://doi.org/10.1145/1645953.1645966

[7] Microsoft. 2023. Bing. https://www.bing.com/?cc=de

[8] OpenAI. 2023. ChatGPT. https://chat.openai.com/auth/login

[9] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. (2018).

[10] Bernardino Romera-Paredes and Philip H. S. Torr. 2015. An Embarrassingly Simple Approach to Zero-Shot Learning. https://doi.org/10.1007/978-3-319-50077-5_2 Series Title: Advances in Computer Vision and Pattern Recognition.

[11] Mesay Samuel, Lars Schmidt-Thieme, D. P. Sharma, Abiot Sinamo, and Abey Bruck. 2022. Offline Handwritten Amharic Character Recognition Using Few-shot Learning. http://arxiv.org/abs/2210.00275 arXiv:2210.00275 [cs].

[12] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y.-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. http://arxiv.org/abs/2208.03188 arXiv:2208.03188 [cs].

[13] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. http://arxiv.org/abs/2201.08239 arXiv:2201.08239 [cs].

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. 30 (2017).