Analyzing User Prompt Quality: Insights From Data

1st Md Asifuzzaman Jishan
Department of Computer Science
Universität Potsdam
Potsdam, Germany
md.jishan@uni-potsdam.de

2nd Mustafa Wasif Allvi
Department of Computer Science
Universität Potsdam
Potsdam, Germany
wasif.allvi@uni-potsdam.de

3rd Md Ataullah Khan Rifat
Department of Computer Science
Universität Potsdam
Potsdam, Germany
khan.rifat@uni-potsdam.de

2024 International Conference on Decision Aid Sciences and Applications (DASA) | 979-8-3503-6910-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/DASA63652.2024.10836452

***Abstract*—This research investigates the quality of user prompts in Large Language Models (LLMs), such as ChatGPT, focusing on the impact of user experience and prompt characteristics on model performance. For this research, we prepared an online survey consisting of questions to collect data and, involving a total of 91 observations, took part in this research over 15 days. The study reveals that month of user experience and prompt satisfaction significantly enhance the ROUGE F1 scores, indicating better alignment between prompts and generated outputs. However, factors such as user background and output satisfaction show no significant impact. Responses were pre-processed and analyzed using statistical packages in Python. The findings underscore the importance of designing longer, well-structured prompts to optimize LLM performance, contributing to improved interactions and outputs in artificial intelligence applications.**

***Index Terms*—Large language models (LLMs), prompt quality user experience, ROUGE F1 score, prompt design, model performance**

I. INTRODUCTION

Prompts are crucial in work and daily life, especially for Large Language Models like GPT-4. However, there is a lack of research and training on prompt engineering, and there is no standard set of rules for writing good prompts. Failure to create good prompts can lead to LLMs finding irrelevant or wrong answers, hindering goal achievement. This paper addresses the need for a formal scheme to adjudicate quality in promptness, as understanding what prompts work better is essential for decision-making and getting things done. Failure to do so may decrease the utility of LLMs and lead to confusion and missed opportunities [1], [2].

Few studies address the criteria and methods used to evaluate prompt quality. There are some other research studies performed on other parts of LLM success, but it is quite explicit that very few examined how the questions were written and how it changed the result. This research should close this gap promptly, as in essence it offers us a new way of thinking to better how people are using the LLM [3]. Without realizing this, a scientist or pro would likely be missing a key aspect that would affect how well the LLMs might work out there in the real world. This paper answers the question:

How do we evaluate the quality of user prompts?

The study used the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) method to measure the ROUGE-L F1 score, focusing on user factors such as educational background, LLM expertise, and satisfaction with prompts. Multiple linear regression analysis was used to establish an association between the ROUGE-L F1 score and user factors. The results showed that user value is a key determinant of the value of prompts. If prompts are pleasing to humans, they perform better on prompts. Researchers can use user feedback data to advance prompt engineering processes, making it a valuable resource for future prompts [4], [5]. This paper contributes additional empirical evidence and theory development toward continuing conversations about effective human-computer interaction, especially those that focus on the support of high-quality prompts. In addition to increasing academic understanding, this research metric for prompt assessment offers useful feedback for the construction of enhanced LLM models and the support of their effective and efficient deployment in applications [6], [7].

This research is important in its immediate findings. Still, it brings out a more important trend—that people are becoming more dependent on LLMs in most aspects of professional and daily life. With LLMs being more and more infused into the decision-making process and its inclusion in various daily tasks, the requirement for effective, prompt engineering becomes critical. This work closed one of the real gaps in the literature and opened further studies that could be undertaken to develop more refined and standardized approaches to prompt engineering. The findings in this study would then make for a great base from which easier, more interactional models of LLMs could be created and their overall usefulness and reliability increased.

The organization of the entire paper is as follows: This paper begins with an introduction that emphasizes the growing importance of effective prompt engineering for Large Language Models (LLMs) and the lack of standardized methods for creating high-quality prompts. Following that, a theoretical background section includes relevant information regarding past research work on this matter. After the theoretical background section, we describe the methodology, followed by results, discussion, and conclusions, which summarize the whole research's outcome. The references utilized for this research can be located in the final section.

II. THEORETICAL BACKGROUND

Large Language Models (LLMs) are AI systems that leverage deep learning to perform tasks like content creation, trans-

lation, and sentiment analysis with minimal task-specific data. Despite their advanced capabilities, little research has focused on how prompt engineering impacts LLM performance. This gap is significant because prompt quality directly affects the accuracy and usefulness of LLM outputs in practical applications. Studies, such as one using GPT-3.5 for job classification, show that even small changes in prompts can significantly enhance performance, especially in zero-shot and few-shot learning. Another study highlights how prompt patterns can improve software engineering processes, helping automate tasks like API generation and code quality improvement. These findings underline the importance of prompt engineering in maximizing the practical utility of LLMs across diverse fields, from job recruitment to software development [8], [9].

Prompt engineering is described as a deliberate approach to crafting prompts that enable Large Language Models (LLMs) to generate more valuable responses. Engineers refine prompts by understanding model behavior or iterating on prompt structures to improve outputs. Small modifications to prompts can significantly influence LLM performance, making the quality of prompt engineering essential for effective real-world application [10].

LLMs also offer educational benefits, helping students of various skill levels. Beginners benefit from structured prompts, while advanced users thrive with open-ended prompts that foster creativity. Teachers can leverage LLMs to provide personalized learning experiences [11]. In AI-assisted coding tools like GitHub Copilot, performance varies with user expertise. Developers who understand how to craft effective prompts perform better. Continuous learning and adapting to AI tools are vital for improving productivity and code quality [12], [13].

While these studies highlight the benefits of Large Language Models (LLMs), they also emphasize a crucial gap in the literature regarding prompt quality assessment. Addressing this gap is essential due to the increasing use of LLMs across various fields. The paper argues that well-crafted prompts, which clearly guide the model, significantly enhance LLM performance. Prompt engineers can use structured patterns and reusable designs to improve code quality, refactoring, and software design. By providing specific instructions in prompts—such as formatting rules, examples, or word limits—engineers can ensure LLMs produce high-quality, modular outputs. These structured prompts not only reduce errors but also streamline software development and other tasks [14], [15].

III. METHODOLOGY

The methodology section relates to all the systematic procedures and techniques used in this research that make the results both reliable and replicable. Research methodology is informed by the thorough investigation of factors impacting the quality of prompts utilized with LLMs - specifically, those related to ChatGPT.

In order to answer the research question: 'How do we evaluate the quality of user prompts?' This research adopted a quantitative research design via an online survey to collect data regarding user's experiences and satisfaction levels with LLMs. Survey methods were chosen because they can collect a huge quantity of data from very different participants in a relatively short period of time, so this fits this research objective of understanding how users interact with LLMs.

To collect data for this research, conducted an online survey using the survey service umfragenup, which was selected as the survey platform by the University of Potsdam. The survey link stayed accessible to the public for a duration of 15 days (from 17.06.2024 to 02.07.2024), at which point we stopped accumulating responses.

The survey aims to assess user interaction with Large Language Models (LLMs), such as ChatGPT, by collecting responses on several key factors. Participants are asked about their experience with LLMs, including the duration of usage and their field of work or study. Additionally, the survey evaluates user satisfaction with creating prompts and the effectiveness of LLM-generated results using a 5-point scale. Participants are also tasked with constructing a prompt to guide the LLM in developing a comprehensive plan for learning a new skill. This task emphasizes the importance of prompt design in LLM performance, with structured prompts leading to better, more accurate results. The survey highlights the variability in user experience based on prompt quality and seeks to identify patterns that can enhance LLM output across different domains, including education, software development, and general task automation. For the convenience of the participants, the authors only considered ChatGPT as an LLM model for this research. The responses were pre-processed so that the analysis was free from any inaccuracies and inconsistencies. Pre-processing followed these steps:

- **Data Cleaning**: Omission of incomplete responses - 3 missing values were found and dropped.
- **Data Formatting**: Standardization of the format of the responses, especially open-ended questions and Likert scale responses.
- **Coding and Categorization**: Quantitative data were numerically coded for statistical analysis.

The dataset had a total of 91 complete responses from individuals at all levels of experience in using LLMs like ChatGPT. There are four major variables within the dataset: Background, Month of Experience, Prompt Satisfaction, and Output Satisfaction. The Background variable is coded numerically for the representation of various professional fields like Technology/IT, Business, and Arts and Humanities. Month of Experience also adds into the primary depiction of how long each subject has had LLMs; from 1 to over 24 months. Satisfaction with both the prompt and the output are measured using a 5-point Likert scale where 1 represents 'Very Low' and 5 represents 'Very High'.

For instance, a respondent with 10 months of experience in Arts and Humanities rated their prompt satisfaction as 4, and output satisfaction as 4. Their prompt was a request for some simple, holistic plan on how to begin picking up a new skill, like an instrument or a language. The ROUGE F1 score for this prompt was 0.2656, indicating medium agreement between the

generated output and the intended result. This dataset has a highly critical ROUGE F1 score as the measure of the quality of any prompt. The value for this ranges between 0, which indicates no match, and 1, which is a perfect match. The differences across ROUGE scores for this dataset reflect how effective prompts are that were created by users with different levels of experience and different levels of satisfaction.

The ROUGE score is a collection of metrics utilized to assess the quality of summaries by comparing them to one or more reference summaries. In the ROUGE score, there are several variants which are ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S, where ROUGE-N quantifies the similarity between n-grams (usually unigrams, bigrams and trigrams). ROUGE-L quantifies the longest common subsequence between the candidate summary and the reference summary, ROUGE-W calculates a weighted score based on the longest common subsequence, taking into consideration the length of matching subsequences, and ROUGE-S measures the similarity between skip-bigrams, which are pairs of words in their original sentence order, allowing for gaps between the words [4]. In this research, ROUGE-L was used because it is proven to be much more effective at capturing the sequential relations of words, key text components, in the quality assessment of prompts designed for eliciting coherent and relevant outputs from LLMs.

In order to find out the relationship between the characteristics of users and prompt quality, as measured by ROUGE-L F1 scores, a multiple linear regression model was used [16]. Therefore, it was appropriate to achieve the research objective of identifying some key factors that influence prompt quality. According to the dataset, the multiple linear regression model is:

$$\textit{Rouge F1} = \beta_0 + \beta_1 \times \text{Background} + \beta_2 \times \text{Experience} + \beta_3 \times \text{Prompt Satisfaction} + \beta_4 \times \text{Output Satisfaction} \quad (1)$$

Where:

- Rouge F1 is the dependent variable, representing the quality of prompts.
- β_0 is the intercept of the regression model.
- $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients for the independent variables.
- Backgroud refers to the primary feild of study of user.
- Month of Experience quantifies the participant's exeriience with LLMs.
- Prompt Satisfaction measures how satisfied the user is with their prompt creation.
- Output Satifaction measures how satisfied the user is with the LLM's output.

This regression model was fitted to the total dataset, and the results were used to make interpretations for how each of the independent variables influences prompt quality.

IV. RESULTS

This section represents the outcome of the research results. The distributions of various factors involved in this experiment are represented in Figure 1. Most participants come from a background in information science and technology or IT, labelled as Background 1 in the research.

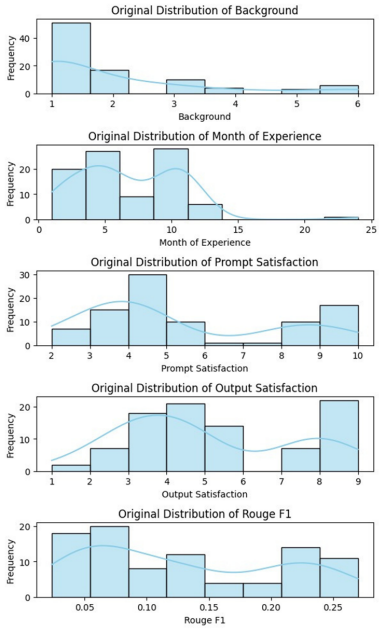

Fig. 1. Original distributions of Background, Month of Experience, Prompt Satisfaction, Output Satisfaction, and Rouge F1.

The frequency of participants decreases as the background of the user's changes. Moreover, the distribution of the Month of Experience is skewed towards the lower end, indicating most participants have fewer months of experience as an LLM model user, with a notable peak around 5 to 10 months. In addition, in the Prompt Satisfaction and Output Satisfaction features both display a somewhat normal distribution with peaks around the satisfaction levels of 4 and 5, respectively. The distribution of Rouge F1 scores indicates that most scores cluster around 0.1 to 0.15.

The square root transformed distributions of the variables involved in this experiment are represented in Figure 2. The square root transformation minimizes the influence of large values in data, helping to stabilize variance and make the distribution more normal, which is beneficial for accurate statistical analysis. Transforming the data using the square root method is crucial because it can help us stabilize variance and make the data more normally distributed, a common assumption in many statistical analyses. The transformed distribution of Background still shows a higher frequency at the lower end, with most participants having a transformed background value close to 1. The Month of Experience distribution is more normally distributed than the original, with most values between 1.1 and 1.4. Similarly, Prompt Satisfaction and Output Satisfaction features now display more symmetric distributions with peaks around 1.2 and 1.15, respectively. Lastly, the transformed Rouge F1 scores have become more evenly spread, clustering around 0.7 to 0.8, suggesting improved data normalization, which always helps us get better accuracy.

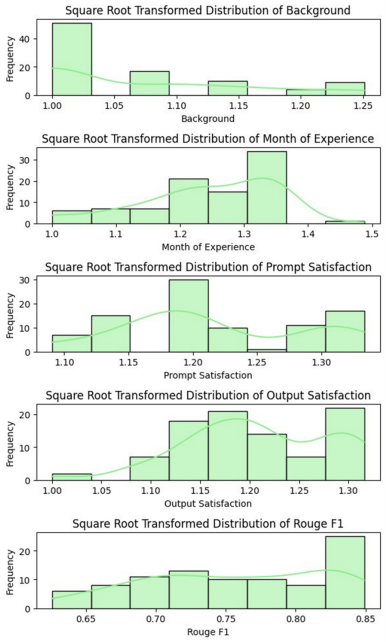

Fig. 2. Square root transformed distributions of Background, Month of Experience, Prompt Satisfaction, Output Satisfaction, and Rouge F1.

The Ordinary Least Squares (OLS) regression results are represented in Table 1. The results indicate that the Month of Experience feature and Prompt Satisfaction feature significantly predict Rouge F1 scores. Specifically, the Month of Experience feature has a coefficient of 0.0289 (p = 0.042), and the Prompt Satisfaction feature has a coefficient of 0.1265 (p ¡ 0.001), both showing positive and statistically significant effects. This suggests that for each additional month of experience, the Rouge F1 score is expected to increase by 0.0289, and for each unit increase in prompt satisfaction, the Rouge F1 score is expected to increase by 0.1265. Here, consider the threshold for p-value as 0.05, which means a 5% significance level. The null hypothesis is rejected for these variables, indicating that they significantly contribute to the model.

The constant term (-0.0374, p = 0.412), the Background feature (0.0024, p = 0.894), and the Output Satisfaction feature (0.0095, p = 0.753) do not have significant impacts on Rouge F1 scores, as their p-values are above the 0.05 threshold. In these cases, the null hypothesis is supported, indicating that these variables do not significantly predict the Rouge F1 scores. In OLS regression, the constant term, also called the intercept, represents the expected value of the dependent variable (specifically, Rouge F1 scores) when all independent variables (such as Month of Experience, Prompt Satisfaction, etc.) are equal to zero. It essentially indicates the baseline level of the dependent variable when there are no contributions from the other variables in the model. The model obtained 61.6% of the variance (R-squared = 0.616), so we can say that 61.6% of the variability in the dependent variable can be explained by the independent variables in the model. These findings highlight the importance of experience and prompt satisfaction in improving performance outcomes.

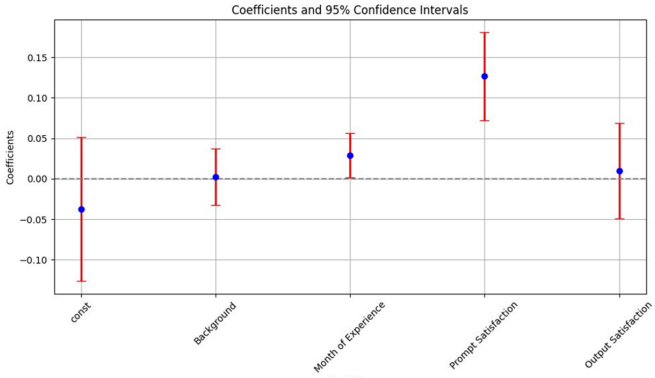

Fig. 3. Coefficients and 95% Confidence Intervals for const, Background, Month of Experience, Prompt Satisfaction, and Output Satisfaction.

Figure 3 represents the coefficients and 95% confidence intervals for the constant term, Background, Month of Experience, Prompt Satisfaction, and Output Satisfaction features. The figure shows that the coefficient for the constant term is slightly negative, indicating a minimal baseline effect. Moreover, Background features show a near-zero coefficient with a wide confidence interval crossing zero, which suggests that there is no significant impact. To extend, the Month of Experience feature has a positive and significant coefficient, indicating that increased experience positively affects the outcome. On the other hand, the Prompt Satisfaction feature has the highest positive coefficient with a confidence interval above zero, confirming it has a strong positive and significant effect. In addition, the Output Satisfaction features coefficient is positive but not statistically significant, as its confidence interval of zero lies between the lower limit and upper limit.

V. DISCUSSION

This research examines factors affecting prompt quality in Large Language Models (LLMs), focusing on user experience and satisfaction. OLS regression results show that each additional month of experience increases ROUGE-L F1 scores by 0.0289, and each unit rise in prompt satisfaction adds 0.1265. Both factors are statistically significant, suggesting that more experience and satisfaction lead to better prompt quality, consistent with theories on expertise and self-efficacy.

The model explains 61.6% of the variation in prompt quality, with diagnostics confirming its reliability and no overfitting. However, educational background and satisfaction with LLM output did not significantly affect prompt quality, indicating that external factors may influence user satisfaction.

The findings suggest that more experienced users craft better prompts, likely due to a deeper understanding of LLMs. While the study is limited by its small sample size and focus on text-based prompts, it opens avenues for exploring cognitive factors and task complexity in future research. These insights emphasize the need for prompt-writing training to improve LLM interactions, with the potential for developing tools and

TABLE I
OLS REGRESSION RESULTS FOR ROUGE F1 AGAINST CONST, BACKGROUND, MONTH OF EXPERIENCE, PROMPT SATISFACTION, AND OUTPUT SATISFACTION

OLS Regression Results						
Dep. Variable:	Rouge F1		**R-squared:**	0.616		
Model:	OLS		**Adj. R-squared:**	0.598		
Method:	Least Squares		**F-statistic:**	34.51		
Date:	Thu, 04 Aug 2024		**Prob (F-statistic):**	3.63e-17		
Time:	21:12:49		**Log-Likelihood:**	114.76		
No. Observations:	91		**AIC:**	-219.5		
Df Residuals:	86		**BIC:**	-207.0		
Df Model:	4		**Covariance Type:**	nonrobust		
	Coef	**Std Error**	**t**	**P** $> \|t\|$	**[0.025**	**0.975]**
Const	-0.0374	0.045	-0.824	0.412	-0.128	0.053
Background	0.0024	0.018	0.134	0.894	-0.033	0.038
Month of Experience	0.0289	0.014	2.063	0.042	0.001	0.057
Prompt Satisfaction	0.1265	0.028	4.552	0.000	0.071	0.182
Output Satisfaction	0.0095	0.030	0.316	0.753	-0.050	0.069
Omnibus:	1.746		**Durbin-Watson:**	2.234		
Prob(Omnibus):	0.418		**Jarque-Bera (JB):**	1.319		
Skew:	0.288		**Prob(JB):**	0.517		
Kurtosis:	3.130		**Cond. No.:**	28.6		

guidelines to assist users, including beginners, in creating effective prompts.

VI. CONCLUSION

According to this study, effective, prompt engineering is crucial for large language models (LLMs) like ChatGPT. We found that user experience and satisfaction improve prompt quality with ROUGE-L F1 scores. This research shows that user experience and prompt satisfaction are essential determinants of LLM performance and that confident users can generate high-quality prompts. As user experience grows, users can write more effective prompts by understanding how LLM output works. This study advances prompt engineering, but it has limitations. The small sample size and text-based prompts limit the findings. Future research should expand the dataset and consider different prompts to understand prompt quality better. Factors like cognitive load and task complexity could enhance prompt engineering insights with more variables.

In conclusion, this research is for academic knowledge and offers practical suggestions for prompt design to improve human-computer interactions. LLM users should be trained for prompt writing to improve getting accurate results. The study's findings lay the groundwork for tools and guidelines to help users write effective prompts, achieving more efficient and effective LLM use across domains. Further research on this related topic will be how to connect advanced artificial intelligence with prompt engineering. While this is a constantly changing field, lessons learned from this study might contribute to the establishment of at least some ground for further academic research into and practical application of how to balance the level of user experience with the growing capabilities of Artificial Intelligence systems.

REFERENCES

[1] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, and D. Amodei, "Language models are few-shot learners", *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint*, arXiv:1810.04805, 2018.

[4] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries", *Text Summarization Branches Out*, pp. 74–81, 2004. Available: https://aclanthology.org/W04-1013/

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners", *OpenAI Blog*, vol. 1, no. 8, pp. 9, 2019.

[6] A. Bandura, "Self-efficacy: The exercise of control", Macmillan, 1997.

[7] M. Williams, P. Burnap, and L. Sloan, "Towards an ethical framework for using AI in human decision-making", *AI & Society*, vol. 36, no. 2, pp. 345–357, 2021.

[8] B. Clavié, A. Ciceu, F. Naylor, G. Soulié, and T. Brightwell, "Large language models in the workplace: A case study on prompt engineering for job type classification", *arXiv preprint*, arXiv:2303.07142, 2023.

[9] J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D. C. Schmidt, "ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design", *arXiv preprint*, arXiv:2303.07839, 2023.

[10] O. Fagbohun, R. M. Harrison, and A. Dereventsov, "An empirical categorization of prompting techniques for large language models: A practitioner's guide", *arXiv preprint*, arXiv:2402.14837, 2024.

[11] M. Taras, "The use of tutor feedback and student self-assessment in summative assessment tasks: Towards transparency for students and for tutors", *Assessment & Evaluation in Higher Education*, vol. 26, no. 6, pp. 605–614, 2001.

[12] J. Sweller, "Cognitive load during problem solving: Effects on learning", *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.

[13] B. Yetiştiren, I. Özsoy, M. Ayer, and E. Tüzün, "Evaluating the code quality of AI-assisted code generation tools: An empirical study on GitHub Copilot, Amazon CodeWhisperer, and ChatGPT", *arXiv preprint*, arXiv:2304.10778, 2023.

[14] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, "The role of deliberate practice in the acquisition of expert performance", *Psychological Review*, vol. 100, no. 3, pp. 363–406, 2006.

[15] Z. J. Wang, A. Chakravarthy, D. Munechika, and D. H. Chau, "Wordflow: Social prompt engineering for large language models", *arXiv preprint*, arXiv:2401.14447, 2024.

[16] D. C. Montgomery and G. C. Runger, "Applied statistics and probability for engineers", John Wiley & Sons, 2014.