Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Meta-Prompting as a Solution to Students' Prompt Engineering Difficulties for an Optimized Use of GenAI LLMs in the Context of Education: A Quasi-Experimental Study using Mistral Model

Houda Louatouate[1], and Mohammed Zeriouh[2]
[12]Department of Educational Sciences and Technologies of Teaching, Higher School of Teachers ENS, Moulay Ismail University, Morocco
Corresponding Author: Houda Louatouate1, **E-mail**: h.louatouate@edu.umi.ac.ma

| ABSTRACT
The emergence of prompt engineering as a rising field of GenAI has garnered attention with the purpose to resolve ambiguities accompanying its use. In essence, a perfect input prompt performing well on LLM A might not perform well on LLM B as there is a certain disparity in how each model behaves, along with students' poor prompting competency can make of crafting excellent prompts almost impossible to achieve. To address this problematic issue, the present study sheds light on meta-prompting which can solve students' countless difficulties encountered with forming appropriate prompts, besides disparities in how different LLMs respond to the same prompt. To this end, the study adopts a within-subjects quasi-experimental design, with a sample involving N=50 undergraduate students of the Higher School of Teachers – Moulay Ismail University. For data analysis, the study uses SPSS version 25 for statistical representation of data, and Python code executed on Google Colab coding environment in which the Wilcoxon Signed-Rank Test for paired samples was conducted. Results demonstrated that there is a strong significant difference between students' self-crafted prompts and meta-prompts in terms of specificity, comprehensiveness, logical sequence of instructions, and good structure criteria. The direction of the difference is positive suggesting an increase in the overall rating scores between pre-test and post-test results. The present paper has also proven that students' confidence with prompts significantly increases with the use of meta-prompting technique compared to traditional prompting. Ultimately, the study identifies limitations and offers recommendations orienting future research projects in the field of GenAI and LLMs.

| KEYWORDS

AI, Confidence, GenAI, LLMs, meta-prompting, Mistral, prompt, prompt engineering

| ARTICLE INFORMATION

ACCEPTED: 19 April 2025 **PUBLISHED:** 22 April 2025 **DOI:** 10.32996/jcsts.2025.7.2.21

1. Introduction

The popularity of GenAI systems is steadily rising as individuals from various fields, including the educational one, have been adopting different AI models to achieve different learning objectives. The growing use of these systems has raised many queries that scholars in the literature attempt to answer. One very important factor explaining the wide use of GenAI NLP-powered models is their user-friendly nature, allowing users with different backgrounds and with no expertise in computer science to benefit from their support using natural language. In spite of these qualities, the interaction with these models requires a set of skills including prompt engineering proficiency, mandatory to ensure rewarding human-machine interaction results. In this respect, researchers' mission is to optimize the use of GenAI by tackling key elements related to its use and difficulties encountered, alongside the needed skills for an optimized use, especially for students in educational contexts.

Research questions

1) Is there a significant difference between meta-prompts and self-crafted prompts from students' standpoint?
2) Does students' confidence with prompts increase with the use of meta-prompting technique?

Research hypotheses

Prompts

H_0: There is no significant difference between meta-prompts and self-crafted prompts from students' standpoint.

H_1: There is a significant difference between meta-prompts and self-crafted prompts from students' standpoint.

Students' confidence

H_0: Students' confidence with prompts does not increase with the use of meta-prompting.

H_1: Students' confidence with prompts increases with the use of meta-prompting.

2. Review of the literature

On Large Language Models (LLMs)

To provide a plain language explanation of LLMs, researchers in [1] claim, "Large language models (LLMs) are deep learning models trained to understand and generate natural language.". Large Language Models are systems capable of processing large amounts of information as well as generating content using natural language. They are widely known nowadays for showing outstanding capacities in providing help and support to individuals in various fields, including the education one. In this concern, researchers in [2] state, "LLMs can be applied universally across nearly all areas of learning and professional activity". It is clarified in [2] that the flexibility of LLMs can be advantageous in both educational and professional contexts. In education, LLMs are well-known for the great support they offer to students who seek their assistance for a variety of tasks, as it is the case of chatbots. In this concern, researchers in [3] mention, "The chatbot is an interactive tool designed to provide instant feed-back and personalized learning across multiple devices, making education more engaging and accessible.". However, concerns were raised on the accuracy of information provided by these models, as expressed by researchers in [4] who state, "In many contexts, LLMs have low accuracy. The model can generate correctly formatted information such as references, that do not exist". The issue raised in [4] spots the light on how LLMs in many cases may not provide reliable information, which can raise confusion concerns and promote the widespread of false information, setting the ground for ambiguity and misinterpretations. In the same concern, researchers in [5] claim, "Current LLMs exhibit certain limitations in numerous tasks, notably reasoning and robustness tasks". In addition to information reliability issues, researchers in [5] highlight how LLMs might be imperfect, showing low performance in several tasks. In spite of the mentioned unpleasant sides of LLMs' use mainly in the field of education, we cannot deny the importance of these models in assisting students with different learning tasks. In this matter, it is worth mentioning that the widespread use of AI models has introduced new skills crucially essential to their effective use and which a large number of students lack, including prompt-engineering skills.

On students' AI-Prompting skills

A prompt is a set of instructions crafted for an AI model to generate a specific content. In education, it is to admit that a large number of students lack very basic knowledge and skills in crafting effective prompts, which results in irrelevant or very vague output content. In many cases, students do not recognize the source of low-quality responses generated by LLMs, which triggers feelings of frustration and disappointments among some, while others start doubting the effectiveness of these AI products. In essence, prompting is a decisive element in how well GenAI systems respond. Researchers in [6] note, "The prompt engineering techniques can be used to improve the performance of language models like ChatGPT, allowing them to generate more coherent, relevant, and sophisticated responses to user inputs". Hence, we can grasp the importance of prompting strategies and how they can shape the quality of the model's output results. Students' lack of prompting skills not only predicts the low-quality performance of the LLM in question, but also provides a clear vision on how hard it is for students to achieve the desired results. In this matter, it is highlighted in [7] that, "Gaining an insight into students' prompting strategies and their perception of AI-generated answers is of great interest to develop an understanding of their use of generative AI". Students are not only required to develop good prompting skills to leverage the quality of the generated content, but also to develop their understanding of how these models work. The fact that many students are not even aware of prompt engineering as an emerging sub-field of AI renders the development of prompting skills for students an extremely challenging task. Some potential solutions to overcome prompting difficulties suggest the use of meta-prompting as an advanced technique, which consists of making AI-models instruct themselves by generating the right prompt suitable to the user's desired task and to the model's behaving mechanism, representing the stepping-stone towards achieving high quality output results and a suitable solution for students' AI prompting challenges.

2. Methodology
2.1. Research design

The present paper adopts a within-subjects quasi-experimental pretest-posttest design with a focus on students' evaluations of their self-crafted prompts and Mistral prompts using meta-prompting technique to test the latter's effectiveness as a potential solution to students' repeatedly encountered difficulties, which stem from their lack of prompt engineering skills mandatory to interact efficiently with AI models.

2.2. Sample

The sample of this paper includes N=50 EFL (English as a Foreign Language) undergraduate students of the Higher School of Teachers – Moulay Ismail University, who were included following their agreement to be part of this study.

2.3. Procedure

Students were first asked to individually write a prompt on how to overcome public speaking anxiety during presentations (Randomly chosen request). After 10 min, students were given a pre-test form to evaluate their prompts. The form contained five main questions, four questions measured on a four-point Likert scale aimed to identify how good their prompts are based on 4 different criteria, besides a fifth question measured on a four-point Likert scale to measure their confidence with their self-crafted prompts. The next step was introducing students to meta-prompting technique, which consists of making an AI model instruct itself after understanding the user's desired output. Students were introduced to the selected model 'Mistral' available online. Once they all had access to it, it was explained to students that they need to write a new prompt starting with "You are a prompt engineer, write a prompt which will generate..." followed by the main aim. In our case, the aim was to generate content about 'how to overcome public speaking anxiety during presentations', the new prompt was run in a new Mistral session. Mistral model generated a prompt containing a set of instructions based on the given aim. Students were given a new form with the same questions of the pre-test form, representing the post-test form. Similarly to their self-crafted prompts, students were asked to carefully evaluate the Mistral self-instructing prompt based on the same 4 criteria, and rate their overall confidence with their new prompt using meta-prompting technique.

2.4. Data analysis

Scores of both pre-test and post-test were analyzed using SPSS version 25 for statistical measures, representation of results (Creation of Clustered Bar Charts), and frequencies. The ordinal non-normally distributed nature of data suggested the use of Wilcoxon Signed-Rank Test, which was conducted using Python code executed on a coding space.

3. Results
3.1. Self-crafted prompts versus meta-prompts

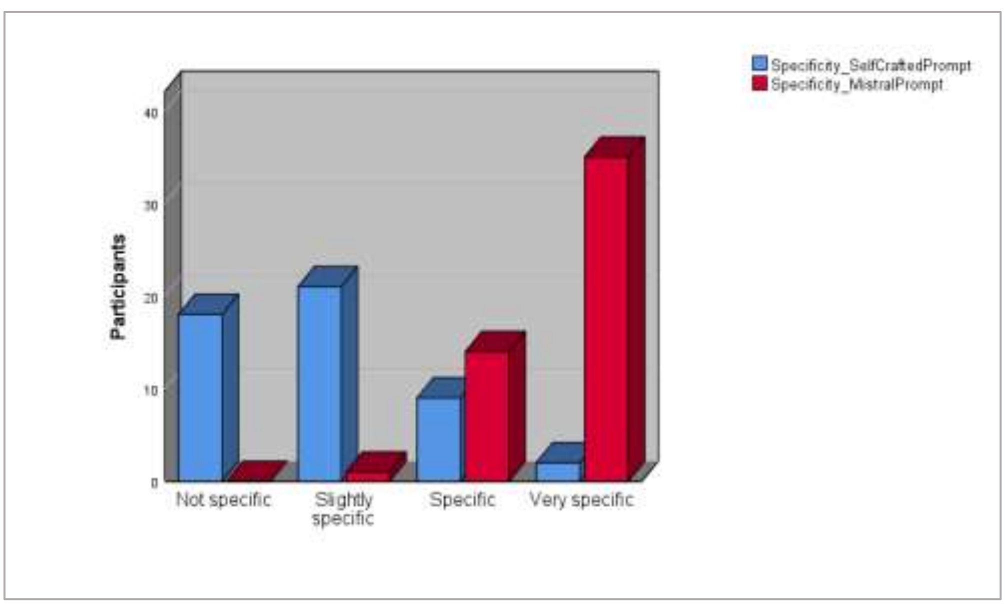

Figure 1. Pre-test and post-test results of prompts' specificity

As shown in Figure 1, 36% (18 participants) mentioned that their individually crafted prompt was not specific, 42% described it as slightly specific, 18% described it as specific, whereas 2% viewed it as very specific. For the meta-prompt, 70% claimed that it was very specific, 28% reported that it was specific, whereas only 2% reported that it was slightly specific. No one reported that their Mistral meta-prompt was not specific.

The paired Wilcoxon Signed-Rank Test, conducted using Python code, has shown a statistically significant difference between scores of the pre-test (median=1) and post-test (median=3) prompts' specificity, with $W=7.00$, $|Z|= 5.93$, and $p<0.001$ ($p = 0.000000003$). Additionally, statistics confirm the positive direction of the difference, suggesting an improvement in the specificity of prompts in post-test results of meta-prompting. Results of the effect size are as follows:

$$r = |Z| / \sqrt{N}$$

$$r = 5.93 / \sqrt{50}$$

$$r \approx 0.84$$

Results suggest that there is a large effect size with r > 0,5. Hence, we conclude with a strong rejection of the null hypothesis in this regard.

```
import numpy as np
from scipy.stats import wilcoxon

# Pre-test and Post-test results
pre_test = np.array([
    2, 1, 3, 1, 0, 1, 0, 1, 0, 1,
    2, 1, 1, 1, 0, 2, 0, 3, 0, 0,
    0, 0, 1, 1, 1, 0, 1, 2, 1, 2,
    1, 1, 2, 1, 1, 2, 1, 1, 2, 0,
    1, 0, 1, 2, 0, 0, 0, 0, 0, 0
])

post_test = np.array([
    2, 3, 2, 3, 3, 3, 3, 3, 2, 3,
    3, 3, 2, 3, 3, 3, 1, 3, 3, 3,
    2, 2, 3, 2, 3, 3, 2, 3, 3, 3,
    2, 3, 2, 3, 3, 2, 3, 3, 3, 2,
    3, 3, 2, 3, 3, 3, 2, 3, 3, 3
])

# Wilcoxon Signed-Rank Test (with exact p-value if possible)
statistic, p_value = wilcoxon(pre_test, post_test, alternative='two-sided', mode='exact')

print(f"Wilcoxon Signed-Rank Test statistic: {statistic}")
print(f"Exact p-value: {p_value}")
```

```
Wilcoxon Signed-Rank Test statistic: 7.0
Exact p-value: 3.098086645127191e-09
```

Example 1. Extracting the exact p value by conducting Wilcoxon Signed-Rank Test using Python code

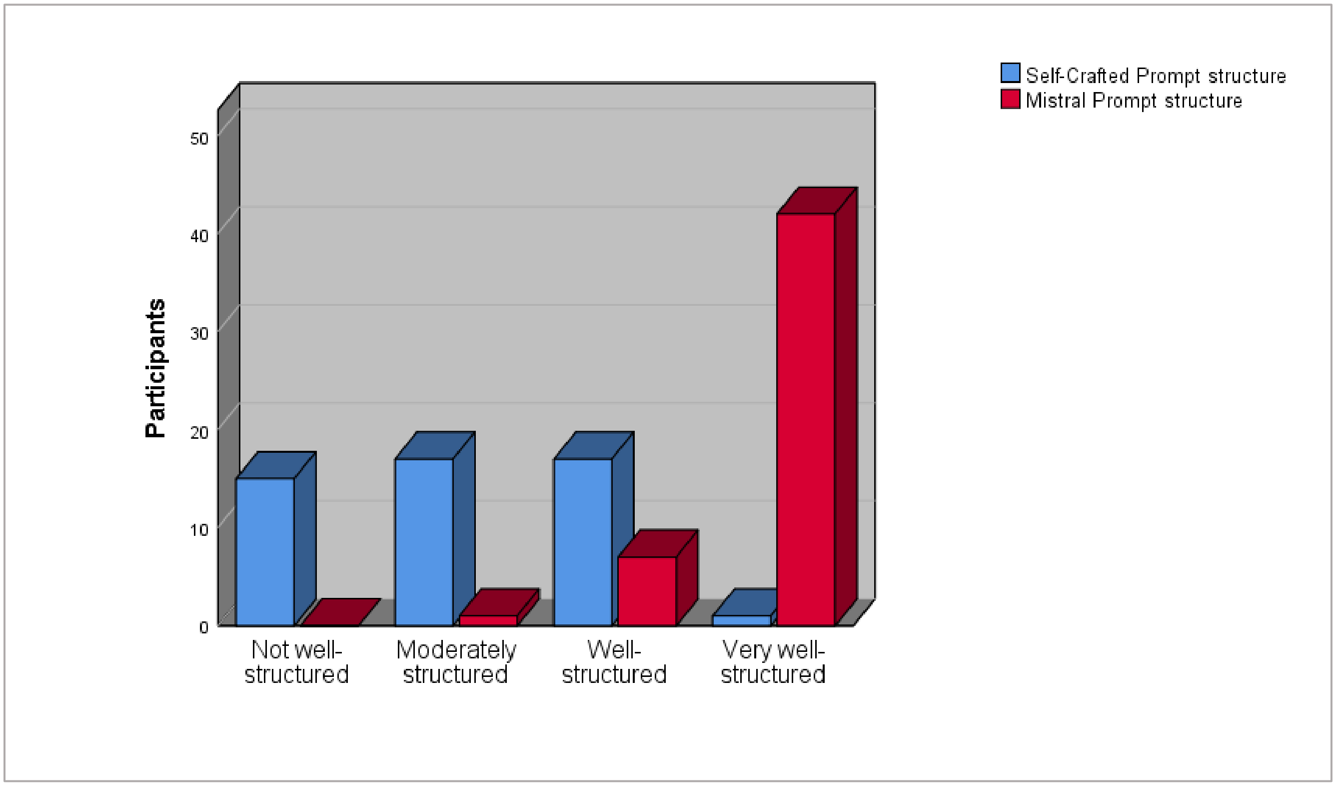

Figure 2. Pre-test and post-test prompts' structure

As shown in Figure 2, 30% of students (15 participants) claimed that their self-crafted prompt was not well-structured, 34% claimed that it was moderately structured, 34% went for well-structured, only 2% described it as very well-structured. For Mistral model meta-prompt, 84% of students reported that it was very well-structured, 14% claimed that it was well-structured, 2% described it as moderately structured. No one reported that the meta-prompt was not-well structured.

Using Google Colab coding environment, Python code was executed to run the paired Wilcoxon Signed-Rank test which has revealed a strong significant difference between pre-test (median=1) and posttest (median=3) results of how well-structured self-crafted and meta-prompts are, with $W= 0.00$, $|Z|= 5.98$, $p < 0.001$ ($p=0.000000002$). Statistical evidence suggests that the difference between post-test and pre-test results is in a positive direction. Results of the effect size are as follows:

$$r = |Z| / \sqrt{N}$$

$$r = 5.98 / \sqrt{50}$$

$$r \approx 0.85$$

Results suggest that there is a large effect size with $r > 0,5$. We conclude with a strong rejection of the null hypothesis in this regard.

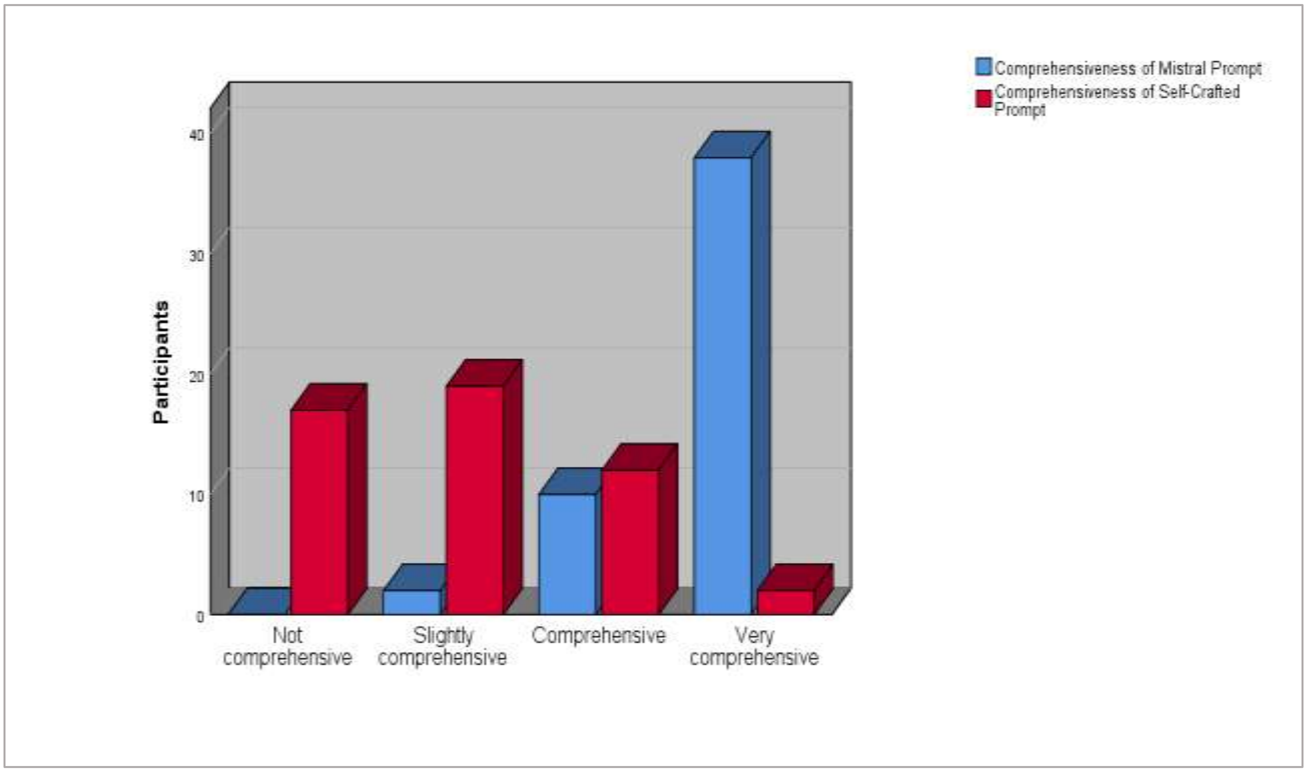

Figure 3. Pre-test and post-test prompts' comprehensiveness

As shown in Figure 3, 34% of participants (17 students) viewed that their self-crafted prompt did not comprehensively cover all details crucial to the desired task, 38% described it as slightly comprehensive, 24% went for comprehensive, whereas only 4% claimed the self-crafted prompt to be very comprehensive. In contrast, 76% of participants claimed that the meta-prompt was very comprehensive, 20% of participants described it as comprehensive, whereas 4% reported that it was slightly comprehensive. No one reported that the prompt they got using Mistral was not comprehensive.

Statistical evidence of Wilcoxon Signed-Rank test revealed a high significant difference between pre-test (median=1) and post-test (median=3) results of comprehensiveness which aimed to measure how comprehensively prompts encompass all necessary details crucial to the desired task, with $W=16.0$, $|Z|= 5.93$, and $p < 0.001$ ($p=0.000000003$). Results of the effect size are as follows:

$$r = |Z| / \sqrt{N}$$

$$r = 5.93 / \sqrt{50}$$

$$r \approx 0.84$$

The obtained results suggest that there is a large effect size with $r > 0,5$. We conclude with a strong rejection of the null hypothesis in this regard.

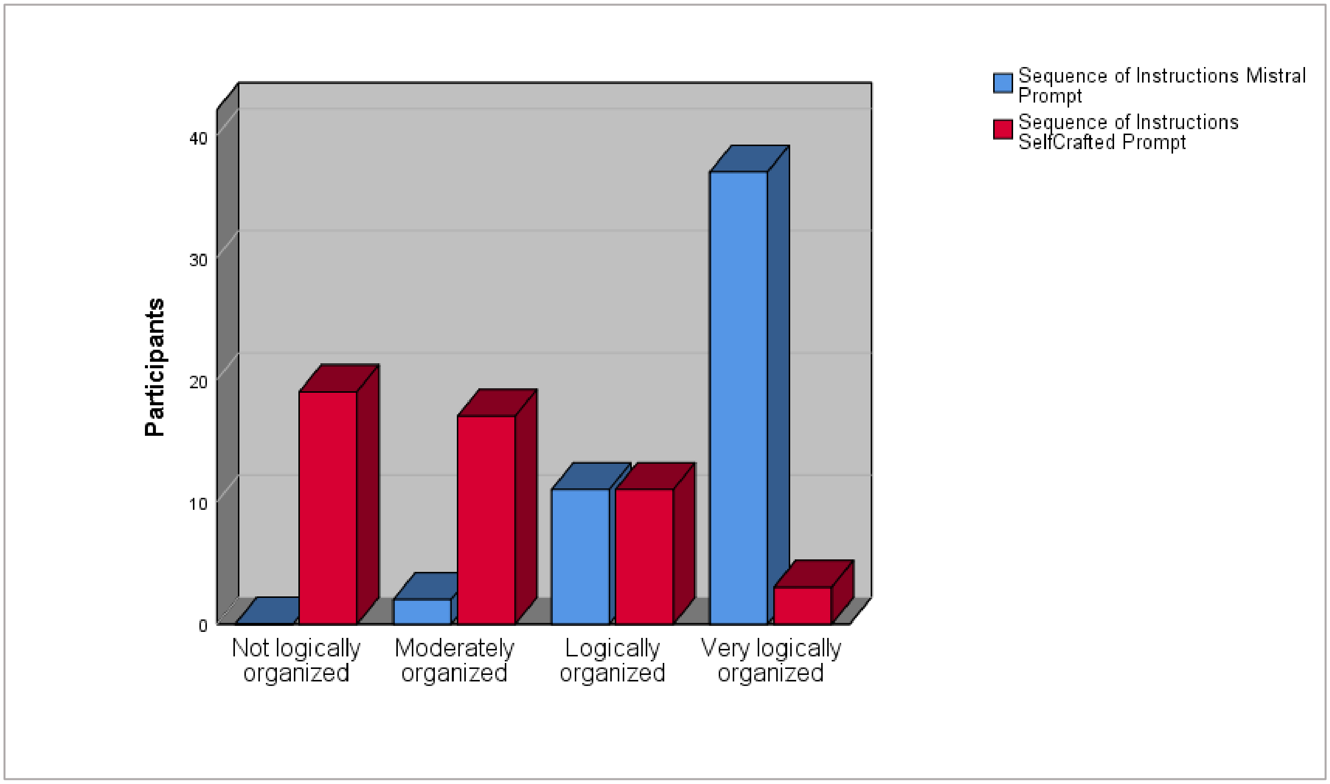

Figure 4. Pre-test and post-test prompts' logical sequence of instructions

As shown in Figure 4, 38% of participants (19 participants) claimed that the self-crafted prompt's instructions were not logically organized, 34% went for moderately organized, 22% went for logically organized, whereas 6% viewed that their prompt instructions were very logically organized. For the meta-prompt, 74% participants claimed that its instructions were very logically organized, 22% went for logically organized, whereas only 4% went for moderately organized. No one reported that the meta-prompt's instructions were not logically organized

We ran the Wilcoxon Signed-Rank test using Python. Results have proven the significant difference between pre-test (median=1) and post-test (median =3) results, with $W=7.0$, $|Z| = 6.02$, and $p < 0.001$ ($p = 0.000000005$). Results of the effect size are as follows:

$$r = |Z| / \sqrt{N}$$

$$r = 6.02 / \sqrt{50}$$

$$r \approx 0.85$$

Statistical results suggest that there is a large effect size with $r > 0{,}5$. Statistical evidence suggests a strong rejection of the null hypothesis in this regard.

3.2. Students' confidence with prompts

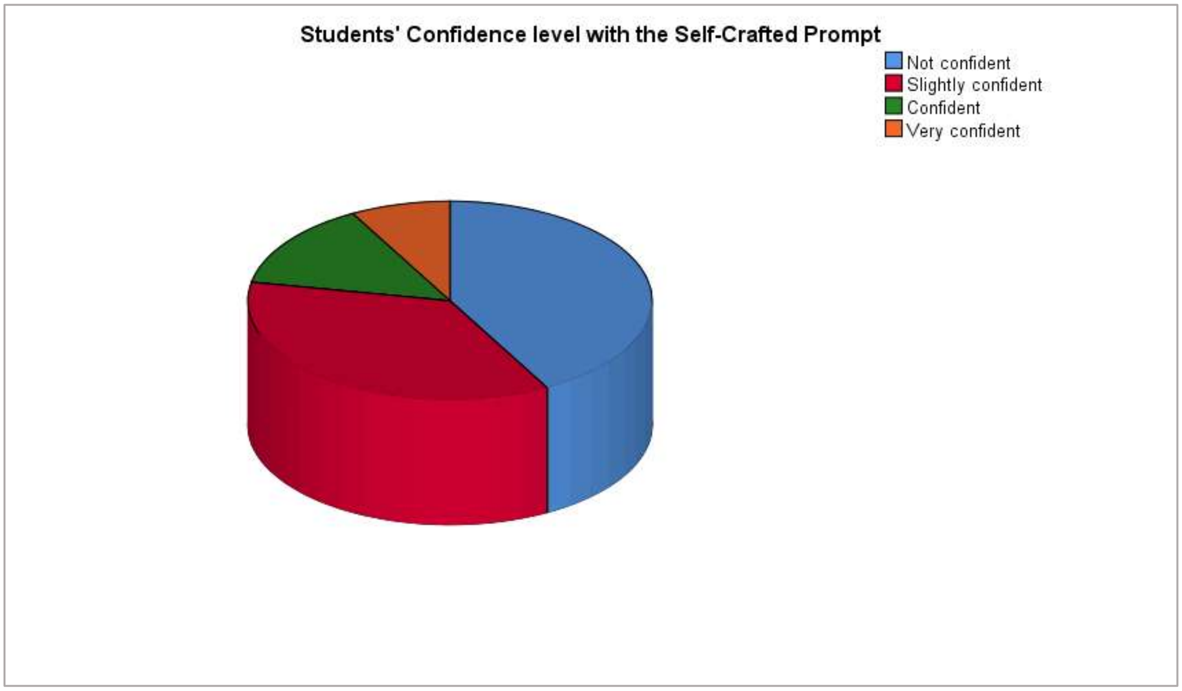

Figure 5A. Pre-test results of students' confidence with self-crafted prompts

As shown in Figure 5A, a large number of participants did not feel confident with their prompts during pre-test results. 42% of participants expressed that they were not confident, representing the majority of participants in the pre-test, 36% reported that they were slightly confident, 14% claimed that they were confident, and only 8% expressed their high overall confidence with the self-formed prompt.

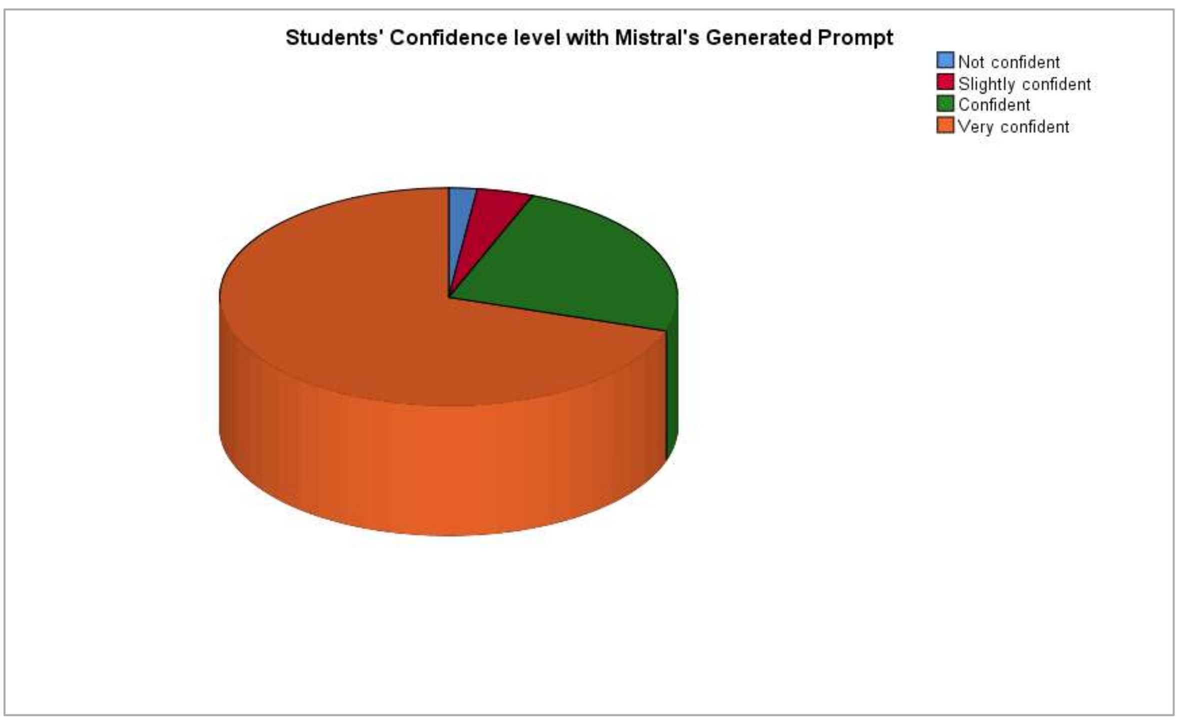

Figure 5B. Post-test results of students' confidence with prompts using meta-prompting technique

Contrary to pre-test results, Figure 5B shows that a majority of 70% of the participants were very confident with the meta-prompts, 24% expressed that they felt confident with meta-prompts, 4% and 2% went for slightly 'confident' and 'not confident' options.

Statistical evidence suggests that there is a strong significant difference between pre-test (median=1) and post-test (median=3) results of students' confidence with self-crafted and meta-prompts, with $W = 13.00$, $|Z| = 5.89$, $p < 0.001$ ($p = 0.00000001$). Results of the effect size are as follows:

$$r = |Z| / \sqrt{N}$$

$$r = 5.89 / \sqrt{50}$$

$$r \approx 0.83$$

Results suggest that there is a large effect size with $r > 0{,}5$. We conclude with a strong rejection of the null hypothesis in this regard.

4. Discussion

Chatbot models have become a central element in today's education due to their crucial role in supporting students as illustrated in [8], "They provide tutoring, personalized learning support, research assistance, and interactive activities aligned with course material, enhancing student engagement.". Other pros associated with AI tools' use have been stated in [9] as the study shows that, "AI tools are valued for their ability to provide concise summaries, answer specific questions, and generate explanations that help students understand complex concepts.". These findings from [8] and [9] partly explain students' preference of these tools in education. Hence, it is very important to make our students equipped with necessary techniques to optimize these tools' performance.

The present study aimed to shed light on meta-prompting which consists of making an AI model generate a set of instructions for itself, suitably adequate to achieve the output results the user (student in our case) aims for after processing his need. Meta-prompting starts by assigning the AI model the role of a prompt engineer and requesting it to generate a prompt (request) based on the description of the output objective, which can operate as a solution to different challenges students face with prompt engineering due to their lack of very basic skills adequate to craft effective prompts. Another issue the paper aimed to address, is the disparity in how different LLMs respond to the same prompt, making it complicated to predict the feasibility of well-crafted prompts in spite of their quality.

- **Specificity**

As suggested, results have shown that meta-prompting using Mistral Model helps students get more specific prompts compared to their self-formed ones. It is worth mentioning that very specific input instructions represent the gateway to achieving accurate output content as AI models need to well-understand the results that the human user is wishing for. According to our findings, one of the main hurdles students encounter in their attempt to well interact with AI models is the lack of specificity in their requests, which logically would lead to ambiguous or vague answers produced by Mistral model or any AI-powered model.

- **Structure**

It was also found that the generated prompts using meta-prompting technique present instructions in a very well-structured manner compared to human self-crafted prompts. It is to be mentioned that well-structured input instructions ensure a highly relevant and clear output result as they reduce risks of misunderstanding and avoid broad answers that AI models might generate in consequence, saving both time and energy. This finding not only introduces us, educators and researchers, to the robust strengths of prompts produced using meta-prompting technique, but also to another aspect of students' lack of brilliant prompt engineering competency, which can lay behind incoherent and poor AI-produced responses.

- **Comprehensiveness**

In terms of prompts' comprehensiveness, meta-prompting technique was found to provide better prompts which comprehensively cover all important details crucially essential to the well understanding of the desired task based on the explained need. It is to clarify that LLMs should have a clear understanding of the task before performing it. Hence, providing all necessary information and details is a fundamental element to ensure the successful achievement of the desired output. Another aspect of students' lack of prompting proficiency unfolds as the findings have shown a large difference in prompts' comprehensiveness between students' self-crafted prompts and meta-prompts. It is to clarify that the lack of details which are crucially essential to the task is the primary reason behind users' struggles. Users with poor prompting skills tend to repeatedly reformulate their requests by adding missing details and eliminating others upon each irrelevant output result, which is unarguably time-consuming, energy draining, and in many cases, frustrating.

- **Logical sequence of instructions**

In terms of prompts' sequence of instructions, it was found that the model's self-instructing prompts were very logically organized. Prompts using meta-prompting technique tend to follow a coherent order, which is essential to maximize Large Language Models' potential. Similar prompts are also challenging for users to craft, especially for those with very little to no competence in prompt engineering in general, and logical organization of ideas in language writing in particular.

- **Students' confidence**

A last construct in the pre-test and post-test forms aimed to measure students' overall confidence with self-crafted and Mistral generated prompts. Results revealed that a great shift in confidence rates has occurred between the two tests as 70% of participants in post-test results have shown higher confidence rates with prompts that they got using meta-prompting technique compared to traditional self-crafted prompts. Our findings which have revealed great disparities in students' confidence rates are not surprising, as meta-prompts' quality was previously reported to be demonstrably superior than traditionally crafted prompts, indirectly affecting students' confidence with prompts' quality.

To conclude this section, findings suggest strong rejections of the null hypotheses. Statistical evidence confirmed that there is a significant difference between prompts created using meta-prompting technique and students' self-crafted prompts, in terms of specificity, structure, comprehensiveness, and logical sequence of instructions. Statistical results have also proven that the direction of the difference is positive suggesting an improvement in quality with:

Post-test prompts' quality results - pretest prompts' quality results > 0

Our findings have also proven that there is a significant difference between students' confidence with prompts with and without using meta-prompting technique. Positive ranks suggest that students' confidence increase with the use of meta-prompting with:

Post-test confidence results – pre-test confidence results > 0

5. Conclusion

The study's mission was to investigate meta-prompting from an educational viewpoint, particularly, as a solution to various barriers students struggle with during their use of GenAI models mainly with prompting. The study highlighted how Large Language Models (LLMs), including Mistral, can instruct themselves after understanding students' need, which can operate as an outstanding strategy to achieve the right customized input prompts, paving the way for effective human-machine interactions in the educational context.

The study compared pre-test and post-test students' self-crafted and meta-prompts evaluation results. Findings demonstrated that students themselves are not able to transform their needs into a well-crafted set of instructions, confirming the importance of similar techniques to facilitates students' use of AI tools for learning. It is to note that meta-prompting operates as a solution not only to students' poor prompt engineering skills, but also to how responses to the same input differ from one LLM to another, making it complicated to predict the generated content's relevance to the user's need based on a given input. Meta-prompting in this case makes LLMs instruct themselves in a way that guarantees the well-understanding of the desired results and to the generation of the most suitable prompt that will make it produce adequate output results depending on its own behaving mechanisms.

Limitations

Limitations of this study include the educational focus on students' pretest and posttest evaluations which might not be fully representative of the prompts' high or low quality from all users' viewpoints.

Recommendations

Students having amazing tools at their disposal but not knowing how to well take advantage of these sophisticated AI solutions can be indeed disappointing and frustrating. Meta-prompting technique can be an outstanding solution to hurdles students encounter due to their poor prompting proficiency. However, this solution is doubtlessly temporary as students need to develop their own robust AI prompting skills. Thus, we highly recommend that educational parties and decision makers consider the incorporation of thorough training sessions and workshops, as it has become an urgent necessity to develop and harness students' digital and AI literacy.

Future research projects should deeply investigate ways in which the repeated use of meta-prompting can help students develop their own individual prompt-engineering competency.

References

[1] Y. Shen *et al.*, "ChatGPT and other large language models are double-edged swords," *Radiology*, vol. 307, no. 2, Jan. 2023, https://doi.org/10.1148/radiol.230163

[2] D. Federiakin, D. Molerov, O. Zlatkin-Troitschanskaia, and A. Maur, "Prompt engineering as a new 21st century skill," *Frontiers in Education*, vol. 9, Nov. 2024, https://doi.org/10.3389/feduc.2024.1366434

[3] K. El Azhari, I. Hilal, N. Daoudi, and R. Ajhoun, "An Integrated AI Specification to Improve Distance Learning," *International Journal of Engineering Pedagogy (iJEP)*, vol. 15, no. 1, pp. 41–55, Jan. 2025, doi: https://doi.org/10.3991/ijep.v15i1.51881.

[4] J. G. Meyer *et al.*, "ChatGPT and large language models in academia: opportunities and challenges," *BioData Mining*, vol. 16, no. 1, Jul. 2023, https://doi.org/10.1186/s13040-023-00339-9

[5] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, Jan. 2024, https://doi.org/10.1145/3641289

[6] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende, "Prompt engineering in large language models," in *Algorithms for intelligent systems*, 2024, pp. 387–402. https://doi.org/10.1007/978-981-99-7962-2_30

[7] S. Tassoti, "Assessment of Students Use of Generative Artificial intelligence: Prompting Strategies and Prompt Engineering in Chemistry education," *Journal of Chemical Education*, vol. 101, no. 6, pp. 2475–2482, May 2024, https://doi.org/10.1021/acs.jchemed.4c00212

[8] A. O. Ajlouni, R. Abu-Shawish, D. M. Silim, and A. H. Ibrahim, "The Academic Intensity Use of Chatbot-Based Artificial Intelligence and Its Relation to Academic Well-Being: A Correlational Study at the University of Jordan," *International Journal of Engineering Pedagogy (iJEP)*, vol. 14, no. 8, pp. 72–87, Dec. 2024, doi: https://doi.org/10.3991/ijep.v14i8.50339.

[9] D. Dobrovská, D. Vaněček, and Yilmaz Ilker Yorulmaz, "Students' Attitudes towards AI in Teaching and Learning," *International Journal of Engineering Pedagogy (iJEP)*, vol. 14, no. 8, pp. 88–106, Dec. 2024, doi: https://doi.org/10.3991/ijep.v14i8.52731.