

Capabilities of Large Language Models in Producing Machine Learning Research

Nikhil Pesati

Received June 05, 2024

Accepted March 02, 2025

Electronic access March 31, 2025

This research explores the potential of Large Language Models (LLMs), particularly ChatGPT, in generating machine learning research. Drawing from the insights of previous studies and experiments, the hypothesis posits that LLMs, when provided with accurate instructions, can effectively produce research components. The study employs a structured approach, inspired by previous methodologies, to evaluate ChatGPT's proficiency in generating idea proposals, background sections, and experimentation strategies. The findings indicate varying levels of success across different research components, with notable challenges observed in accurately representing background information and formulating feasible experimentation strategies. These limitations underscore the importance of cautious reliance on LLMs for critical research tasks and suggest avenues for further exploration to enhance their capabilities.

Keywords: prompt engineering, machine learning, natural language processing, large language models

Introduction

This research project is motivated by the growing body of literature exploring the potential of ChatGPT and other Large Language Models (LLMs) to contribute to academic research. Specifically, the study focuses on evaluating the capacity of LLMs, with a primary emphasis on ChatGPT, in generating high-quality research content within the realm of machine learning. The overarching goal is to investigate whether LLMs can effectively perform research tasks related to machine learning. If the LLMs can carry out these tasks, it would help the computer science world lay the groundwork for potential advancements in the field. Research in the field of natural language processing has shown that Large Language Models (LLMs), such as ChatGPT, are capable of providing well-qualified answers when given sufficient context¹. The model's performance can be optimized by incorporating clarifying questions and contextual information, minimizing the likelihood of ambiguous responses². The theoretical foundation suggests that as long as an LLM is equipped with an ample amount of initial data and complex algorithms, it should demonstrate versatility in executing tasks within the constraints of the provided data³. Given that ChatGPT has been trained on an extensive dataset, it is reasonable to expect that the model possesses the capabilities to successfully generate a research project. Furthermore, insights from the application of ChatGPT in finance research, as highlighted in an article on ScienceDirect by Michael Dowling and his fellow researchers, indicate the model's ability to create articles deemed acceptable for publishing journals by review panels⁴. This suggests

that ChatGPT can contribute meaningfully to the generation of research-related content, particularly in domains such as finance. These foundational studies set the stage for exploring the potential of ChatGPT in research project generation, motivating the hypothesis that the model can effectively produce various components of a research project. The hypothesis suggests that if LLMs, exemplified by ChatGPT, can successfully engage in machine learning research, they have the potential to generate novel ideas and approaches that could contribute to the improvement of their own performance⁵. This process could lead to enhancements in LLM capabilities, fostering a continuous cycle of improvement. By assessing ChatGPT's proficiency in producing quality research within the context of machine learning, this research aims to contribute valuable insights into the model's potential role in advancing knowledge and innovation in this domain. The hypothesis will be tested by using prompt engineering to have ChatGPT produce certain aspects of a research paper and then comparing these sections with those of published articles. These research papers have been split into 6 categories that are easiest to assess for easier evaluation: finance, biology, pollution, cars, animals, and environment.

Methods

To test my hypothesis regarding the effectiveness of ChatGPT in generating research project components, I adopted a 5-step approach inspired by the model presented in Michael Dowling's article on using ChatGPT for finance research⁴. The steps included generating an idea, a background section, a related

works section, an experimentation strategy, and a results section. However, the related works section had no accuracy due to ChatGPT's tendency to fabricate citations. In addition, experimenting with ChatGPT proved to be time-consuming. In order to execute the experiment proposed by ChatGPT, a specific dataset tailored to the experiment's needs would have to be sourced and provided to the model. For example, if the machine learning project aimed to predict fluctuation in the stock value for a company based on news headlines, a dataset that contains historical stock price data would need to be found as well as news headlines related to the company during that period. Finding and feeding the datasets into ChatGPT for each of the 45 trials would be time-consuming. Given the need for extensive data and the associated time constraints, conducting the planned 45 trials was not feasible. Consequently, the evaluation was limited to only three sections: the generated idea, background section, and the proposed experimentation strategy.

In order to ensure that the results ChatGPT produced were as accurate as possible, the prompt was fine-tuned through an iterative process. In order to show the iterative process through which the prompts were created, the evolution of the idea prompt will be used as an example. First, the prompt was designed to be broad, aiming to generate a variety of topics related to machine learning. However, it was too vague and did not provide enough context or guidance, leading to overly general or irrelevant responses. In order to ensure that the topics that were produced could be evaluated with limited human error, the prompt asked for research topics that combined the field of machine learning with fields that I had more knowledge in like biology or economics. Even with this change, the ideas were still vague and had no specific goals. An example of this was shown when the topic produced was to investigate how machine learning can improve business processes. In order to make it as specific as possible, the prompt was fine-tuned using specific scenarios to explicitly draw on the knowledge required to formulate topics. The final prompt that was used in the experiment to generate topics for the research project is the following: "You are a college student who has been tasked with writing a research paper in the field of machine learning and food. Using your knowledge, formulate several potential research topics for your project". This second field in the prompt, which was food in the example, was altered slightly every time it was used in order to diversify the topic ideas that were produced. This iterative process that started with a broad prompt and narrowed down to a specific prompt was used for the generation of all three sections. In the prompts that produced the background section and the experimental section, the idea that was produced by the idea prompt was incorporated. Figure 1, 2, and 3 show examples of the engineered prompt being fed into ChatGPT.

To assess the quality and accuracy of each section, I employed a scale of 1-10, where 1 indicated poor quality and 10 signified high quality. I then asked ChatGPT to rate each section and

provide an explanation for its rating. This process aimed to gain insights into ChatGPT's reasoning and ensure a thorough assessment. Then, I rated the sections as an objective third party. By comparing my ratings with ChatGPT's and analyzing its explanations, I could validate the accuracy of my own assessments, recognizing that my analysis involved a certain level of human bias. This step was crucial in making sure that the generated sections were evaluated with precision.

When conducting the comparison between the ratings assigned by ChatGPT to its research projects and those given by a real person, it was important to account for any potential margin of error in the ratings. The margin of error refers to the degree of variance that could occur due to the subjective nature of evaluating research projects, which is particularly relevant when using AI-driven assessments. When comparing the ratings assigned by ChatGPT to its projects with those assigned to the 45 published research projects, a high level of consistency was observed. Out of the 45 published projects, the ratings assigned by ChatGPT matched my ratings in 42 instances. Also, out of the 45 ChatGPT produced projects, the ratings assigned by ChatGPT matched my ratings in 41 instances. This high match rate suggests that ChatGPT's ability to assess research quality is largely reliable, with a margin of error of only 2.

To visualize and interpret the results, I employed bar graphs. The differences in ratings between ChatGPT and published projects were identified and analyzed, providing quality insights into the reliability and validity of ChatGPT's generated research project components.

Results & Discussion

In order to discern whether there was a statistically significant difference between the accuracy and quality of the sections produced by ChatGPT and the sections in published papers, I calculated the 95 percent confidence intervals for each section. Confidence intervals are a range of values within which a value is expected to be with a certain level of confidence, usually 95%. If the 95 percent confidence intervals of two samples do not overlap, it suggests that there is a statistically significant difference between the two sample means. In order to calculate the confidence intervals for each of the 6 samples, the ratings of the three sections produced by ChatGPT and the ratings of the three sections in the published papers, the average of the 45 ratings for each of the 6 samples was calculated. Then, the standard error of the mean for each sample was calculated and multiplied by 1.96 to get the margin of error. The margin of error was subtracted from the mean of the ratings in the specific sample to get the lower bound of the 95 percent confidence interval and it was added to the mean to get the upper bound of the interval. This 95 percent confidence interval model relies on the assumption that the data is normally distributed and randomly selected, which is true as I randomly picked published

```

model = 'gpt-3.5-turbo'
ideal_prompt = """
    You are a college student who has been tasked with writing a research paper in the field of machine learning and food.
    Using your knowledge, formulate several potential research topics for your project."""
response = generate_newer_gpt_model_response(model, ideal_prompt, TEMP = 1)
print(response)

```

1. Analyzing the impact of machine learning algorithms on food recommendation systems and their ability to personalize food choices based on individual preferences and dietary restrictions.
2. Investigating the use of machine learning in predicting food spoilage and reducing food waste in the supply chain.
3. Exploring how machine learning techniques can be used to optimize food production processes, such as crop management, irrigation scheduling, and yield forecasting.
4. Examining the potential of machine learning in designing personalized meal plans and dietary interventions for improved nutrition and health outcomes.
5. Assessing the effectiveness of machine learning in predicting foodborne illnesses and enhancing food safety measures in the food industry.
6. Investigating the use of machine learning in analyzing consumer food preferences and trends to improve marketing strategies and product development in the food industry.

Fig. 1 Prompt for the Idea Generation as well as the output of the prompt. This snapshot shows the code and the prompt used to get ChatGPT to produce the ideas as well as the sample output of the instruction.

```

model = 'gpt-3.5-turbo'
ideal_prompt = """
    You are a college student who has been tasked with writing a research paper in the field of machine learning and food.
    Using your knowledge, formulate several potential research topics for your project."""
response = generate_newer_gpt_model_response(model, ideal_prompt, TEMP = 1)
print(response)

```

1. Analyzing the impact of machine learning algorithms on food recommendation systems and their ability to personalize food choices based on individual preferences and dietary restrictions.
2. Investigating the use of machine learning in predicting food spoilage and reducing food waste in the supply chain.
3. Exploring how machine learning techniques can be used to optimize food production processes, such as crop management, irrigation scheduling, and yield forecasting.
4. Examining the potential of machine learning in designing personalized meal plans and dietary interventions for improved nutrition and health outcomes.
5. Assessing the effectiveness of machine learning in predicting foodborne illnesses and enhancing food safety measures in the food industry.
6. Investigating the use of machine learning in analyzing consumer food preferences and trends to improve marketing strategies and product development in the food industry.

Fig. 2 Prompt for the Background Section generation. This snapshot shows the code and the prompt used to get ChatGPT to produce the background section of the topic chosen by ChatGPT.

```

model = 'gpt-3.5-turbo'
ideal_prompt = """
    You are a college student who has been tasked with writing a research paper in the field of machine learning and food.
    Using your knowledge, formulate several potential research topics for your project."""
response = generate_newer_gpt_model_response(model, ideal_prompt, TEMP = 1)
print(response)

```

1. Analyzing the impact of machine learning algorithms on food recommendation systems and their ability to personalize food choices based on individual preferences and dietary restrictions.
2. Investigating the use of machine learning in predicting food spoilage and reducing food waste in the supply chain.
3. Exploring how machine learning techniques can be used to optimize food production processes, such as crop management, irrigation scheduling, and yield forecasting.
4. Examining the potential of machine learning in designing personalized meal plans and dietary interventions for improved nutrition and health outcomes.
5. Assessing the effectiveness of machine learning in predicting foodborne illnesses and enhancing food safety measures in the food industry.
6. Investigating the use of machine learning in analyzing consumer food preferences and trends to improve marketing strategies and product development in the food industry.

Fig. 3 Prompt for the Experimental strategy generation. This snapshot shows the code and the prompt used to get ChatGPT to produce the experimental section.

papers from the population of machine learning papers.

The non-overlapping 95 percent confidence intervals for accuracy ratings between ChatGPT’s background sections and those from published works were pivotal in rejecting the hypothesis (see Figure 5). The upper bound of ChatGPT’s confidence interval was 7.6, while the lower bound of the confidence interval for real research projects was 7.7. This lack of overlap signifies a significant difference in the accuracy ratings for background sections produced by ChatGPT and those created by human researchers. A primary factor contributing to the lower accuracy ratings for ChatGPT’s background sections is the phenomenon known as “AI hallucination.” This refers to instances where ChatGPT generates information that is incorrect, misleading, or not grounded in factual data. These hallucinations can result in the inclusion of erroneous details or misinterpretations within the background sections, which significantly undermines the accuracy and reliability of its content. Also, human-generated content is typically subjected to rigorous peer review and fact-checking processes, further ensuring its accuracy.

The evidence indicates that ChatGPT struggles to generate an accurate background section, undermining its ability to represent foundational information in a manner comparable to real

researchers.

Similar to background sections, the non-overlapping 95 percent confidence intervals for experimentation strategies further solidify the rejection of the hypothesis (see Figure 3). This discrepancy highlights ChatGPT’s challenges in creating feasible and effective experimentation strategies. The absence of overlap

These findings highlight the limitations of ChatGPT in specific aspects of research project generation. While the model demonstrated moderate success in generating ideas, its shortcomings in accurately representing background information and formulating effective experimentation strategies are evident. Researchers and practitioners should exercise caution when relying on ChatGPT for these critical components of research projects.

In the study, Figure 4 provides a visual representation of the distribution of idea quality ratings generated by ChatGPT and the published research projects. This figure not only illustrates the consistency of ChatGPT’s output but also highlights the potential for LLMs to contribute significantly to creative and academic endeavors in the future. As LLMs continue to evolve, their ability to generate high-quality, specific ideas could revolutionize fields that rely on ideation and problem-solving.

To assess the overall performance, I calculated the average

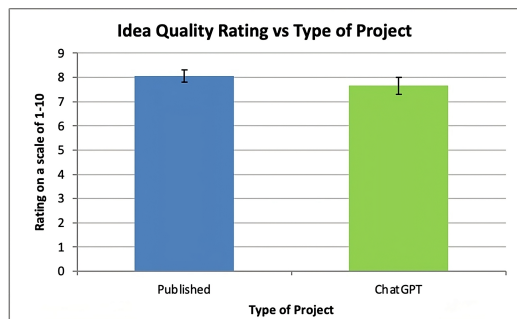


Fig. 4 Idea Quality Rating vs Type of Project. This graph compares the average idea quality rating of a published research project and a ChatGPT-created project on a scale of 1-10. The average idea quality rating for the published projects was 8.07 and the average idea quality rating for the ChatGPT section was 7.65. The 95 percent confidence interval did overlap.

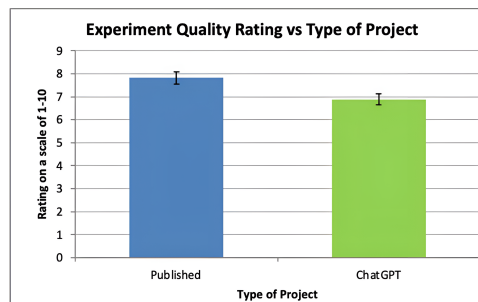


Fig. 6 Experimental Quality Rating vs Type of Project. This graph compares the average experimental section quality rating of a published research project and a ChatGPT-created project on a scale of 1-10. The average experimental quality rating for the published projects was 7.82 and the average experimental quality rating for the ChatGPT section was 6.89. The 95 percent confidence interval did not overlap.

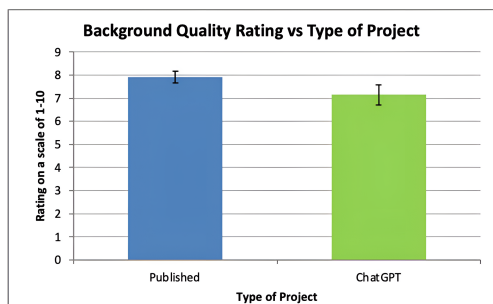


Fig. 5 Background Quality Rating vs Type of Project. This graph compares the average background section quality rating of a published research project and a ChatGPT-created project on a scale of 1-10. The average background quality rating for the published projects was 7.91 and the average background quality rating for the ChatGPT section was 7.15. The 95 percent confidence interval did not overlap. Confidence intervals show that there is a significant difference between the quality of experimentation strategies proposed by ChatGPT and those devised by human researchers.

of all the idea quality ratings generated across the different trials. The idea quality was rated based on its specificity, with higher ratings awarded to ideas that were detailed, actionable, and tailored to the context of the experiment. For instance, an idea that included specific methodologies, clear objectives, and measurable outcomes was rated higher than a vague or overly general suggestion.

Interestingly, ChatGPT consistently produced ideas that scored well on specificity, which contributed to its high average rating. The model's ability to draw from a vast amount of information and synthesize relevant details allows it to generate ideas that are not only creative but also practical and well-suited to the task at hand. This capability suggests that, with further refinement, LLMs like ChatGPT could play a critical role in enhancing the quality and efficient idea generation in various fields.

Conclusion

In conclusion, my hypothesis, stating that ChatGPT is capable of generating accurate and feasible research project components, was proven incorrect on two distinct grounds. These conclusions stem from a statistical analysis of the 95 percent confidence intervals for accuracy ratings in both background sections and experimentation strategies, comparing ChatGPT-generated content with that of real research projects.

Future research could focus on leveraging ChatGPT's creative capabilities to enhance the ideation phase of research projects. This involves exploring methods that allow the model to generate not only innovative ideas but also suggestions for experimental approaches that align more closely with real-world scenarios.

To address the limitations identified in ChatGPT's ability to formulate feasible experimentation strategies, future research could involve a more hands-on approach. This includes having ChatGPT perform the suggested experiments by manually selecting appropriate datasets for evaluation. The results of these experiments could be analyzed to assess the practicality and validity of ChatGPT-generated experimentation strategies.

Improving the accuracy of ChatGPT's responses requires a careful refinement of prompts to minimize hallucinations and inaccuracies. Researchers could experiment with different prompt structures, incorporating more specific guidelines and constraints to guide ChatGPT towards generating more accurate and reliable information. This step aims to enhance the model's understanding of research methods and ensure more precise outputs.

Integrating a human-in-the-loop approach could enhance the reliability of ChatGPT-generated research components. Researchers could actively collaborate with the model, providing additional input, clarifications, or corrections during the generation process. This interactive approach could result in more

accurate and contextually relevant outputs.

Acknowledgments

Thank you to Mr. Spenner, my research teacher, and Alex Ritchie, my mentor for their hard work, skepticism, honesty, and curiosity.

References

- 1 K. Opperman, *How to use ChatGPT: Opportunities and Risks for Researchers*, 2023, Retrieved July 23, 2023, from <https://www.animateyour.science/post/how-to-use-chat-gpt-opportunities-and-risks-for-researchers>.
- 2 P. P. Ray, *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope*, 2023.
- 3 Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja *et al.*, *Opinion paper: "So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy*, 2023.
- 4 M. Dowling and B. Lucey, *ChatGPT for (Finance) research: The bananarama conjecture*, 2023.
- 5 Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan and J. Ba, *Large Language Models are Human-Level Prompt Engineers*, 2023, Retrieved June 25, 2023, from <https://arxiv.org/pdf/2211.01910.pdf>.