INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 11 | NOV - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

A Comparative Analysis of Human-Written vs. AI-Generated Prompts for Task Executions

Nan Wu
Stamford, CT, USA
Chris.wunan88@gmail.com

***Abstract*—This study examines the effectiveness of AI-generated prompts in task execution compared to human-written prompts. AI-generated prompts showed comparable performance to expert-crafted prompts and offered greater flexibility, making prompt engineering more accessible to users. Future work will explore more advanced prompt templates.**

***Keywords*—AI-generated prompts, prompt engineering, task execution, content quality, humanized writing**

Funding Declaration

The author declares that no funding was received for conducting this study or for the preparation of this manuscript.

I. INTRODUCTION

What made AI-powered solution performance a driving factor is the quality of the prompts themselves, serving as one of the main inputs to execute tasks. This has made the evolution of prompt engineering among the fastest-growing areas of interest, and its growing role in AI-generated prompts is an area of increasing interest.

This in itself can let general users master writing more effective prompts without expensive trial or learning of complex prompt techniques. This lowers the barrier for users attempting to harness the powerful capabilities of large language models more effectively. However, how well AI-generated prompts fare against the human-crafted ones-particularly diversified tasks-remains a serious investigation.

This study will make a comparative analysis of human-written and AI-generated prompts on efficiency and accuracy of tasks. The research will target different kinds of tasks: translation, knowledge search and explanation, and humanized writing. By assessing the difference in quality of the results, this research hopes to bring to light how AI-generated prompts can help or hinder overall performance in AI-driven solutions.

II. BACKGROUND AND CONTEXT

Increasing interest in prompt engineering has emerged due to the fast-growing importance of prompt quality in AI-driven applications. Various studies analyse the role of human deliberation over time in properly writing prompts for an AI model, putting the machine toward its generated high-quality output.[1] The strength of human prompts-including their adaptability and awareness of context-and a limitation like variabilities and expertise needed in creating effective human inputs, have underlined by researchers.[2]

Direct empirical studies on AI-generated prompts are few and far between, although, in recent times, progress in artificial intelligence has led to the development of methods to generate prompts, a conception developed to ease interactions with AI. Recent leading examples include Anthropic's workbench feature[3] and the playground feature by OpenAI[4], both resulting from more or less simple inputs and generating AI-written prompts. And this ability could further reduce the barriers for less technologically oriented people by showing them a more consistent way of interacting with models. Because of the sudden pervasion into every field by AI technology, we think that research regarding AI-generated prompts will be of increasingly high priority.

III. METHDOLOGY

A. *Test Tasks*

The design includes three kinds of tasks as described below:

Task 1: Poetry Translation

Task Description: This task involves translating a selected segment of poetry from English to Chinese. Poetry is chosen due to its linguistic complexity, cultural nuance, and inherent ambiguity, making it one of the most challenging forms of text to translate accurately and elegantly.

Evaluation Method: The translated output will be evaluated by a native Chinese speaker. The evaluation will involve comparing the AI-generated translation against an authoritative reference translation, assessing both the fidelity to the original text and the stylistic elegance of the translated output. Special attention will be given to capturing the nuances, emotional tone, and cultural references embedded in the original poetry.

Table I provides the original English text and a sample translation by human expert.

TABLE I. ORGINAL TEXT AND SAMPLE TRANSLATION

Original Text	***Sample Translation***
When the evening is spread out against the sky Like a patient etherized upon a table; Let us go, through certain half-deserted streets, The muttering retreats Of restless nights in one-night cheap hotels And sawdust restaurants with oyster-shells: Streets that follow like a tedious argument Of insidious intent To lead you to an overwhelming question.[5]	正当朝天空慢慢铺展着黄昏 好似病人麻醉在手术桌上； 我们走吧，穿过一些半清冷的街， 那儿休憩的场所正人声喋喋； 有夜夜不宁的下等歇夜旅店和满地蚌壳的铺锯末的饭馆； 街连着街，好象一场讨厌的争议 带着阴险的意图 要把你引向一个重大的问题…… 唉，不要问，"那是什么？" 让我们快点去作客。[6]

Task 2: Historical Knowledge Search and Explanation

Task Description: This task requires providing a concise introduction and an in-depth explanation of a historical event related to World War II, specifically focusing on Japan's 226 Incident. This task tests the ability of the AI to retrieve, synthesize, and present historical information in a coherent and informative manner.

Evaluation Method: The explanation provided by the AI will be validated for accuracy and comprehensiveness by cross-referencing with reliable historical sources[7]. Evaluators will assess whether the explanation captures key details of the event, including its causes, main events, and consequences. The clarity of the explanation, as well as the logical flow of information, will also be evaluated to determine how effectively the prompt guides the AI in constructing a meaningful and informative narrative for general users and provide reasoning behind the score.

Task 3: Humanized Writing Task:

Task Description: This task involves rewriting a text generated by an AI model to enhance its human-like qualities. The goal is to make the text more nuanced, expressive, and natural, mimicking human writing styles more closely.

Evaluation Method: The rewritten output will be evaluated using an AI text detector Zero-GPT[8] to

determine the score to which the text exhibits human-like characteristics. A human evaluator will assess elements such as nuance, natural flow, emotional tone, and contextual relevance.

Fig.1 provides the AI generated text, which all test groups will rewrite based on. The Zero-GPT identified this text as AI written content.

```
One of the hottest topics in recent months has been
the ethical and societal implications of deepfake
technology, which has evolved dramatically due to
advancements in artificial intelligence. Deepfakes,
which involve using AI to generate hyper-realistic
images, videos, or voice clips of people that can
easily deceive viewers, have both amazed and alarmed
the world. Initially seen as a novelty, deepfakes
have now sparked serious concerns around
misinformation, privacy, and trust. On one hand, they
hold immense creative potential in film, media, and
entertainment—enabling actors to "return" to the
screen posthumously or altering scenes without
expensive reshoots. On the other hand, the dark side
of deepfakes is gaining traction. There have been
numerous cases of their use in producing fake news,
identity theft, and revenge pornography, which can
cause emotional distress, reputational damage, and
even impact democratic processes. The sheer ease with
which these tools can be used by virtually anyone
poses significant challenges for social media
platforms, law enforcement, and policymakers.
Recently, the debate has escalated around the need
for stricter regulations and the development of AI
detection tools to counter the rise of deepfake
misuse. Major tech companies are working on methods
to identify and flag deepfakes, while some
governments are considering legislation to penalize
malicious actors. As the technology continues to
evolve, it is clear that society must walk a fine
line between leveraging deepfake tools for creative
and positive uses while actively mitigating the risks
they pose to personal privacy, public safety, and
democratic stability. The conversation about where to
draw the line with deepfake usage will likely
intensify, given the speed at which the technology is
advancing.
```

Fig. 1. *Smaple of AI generated content*

B. Test Gourps

There are three groups of different prompts in this study, each articulated with the GPT-4o model.

Each task will have three repetitions; then, the best result will be picked for further analysis. This approach aims to account for the variability in AI performance and ensure the results truly represent the best possible outcomes for each type of prompt.

Group 1: Simple Human Prompt

This group involves the use of straightforward, manually constructed prompts. These prompts are designed to be direct and unambiguous, providing the AI with clear and concise instructions.

Table II provides prompts used for group 1

TABLE II. PROMPTS USED FOR GROUP1

Task	***Prompt***
1	Translate this poem into Chinese.
2	Tell me about the 226 incident and some of it's shocking details
3	Rewrite this text to make it appear as though it was written by a human, not AI.

Group 2: Complex Human Prompt

This group employs more sophisticated prompts sourced from popular collections available through OpenAI's online GPTs store platform. These agents are designed to incorporate more context and detailed instructions, aiming to guide the AI more precisely.

The prompts used to interact with GPTs agents are the same as these of group 1.

Table III provides GPTs agents used in this group

TABLE III. GPTS USED IN GROUP 3

Task	***GPTs Agent***
1	Translate GPT[9]
2	History[10]
3	AI Humanizer[11]

Group 3: AI-Generated Prompt)

This group utilizes prompts autonomously generated by GPT-4o based on original prompt embedded. Once user input a trigger prompt, the model is directed to ask follow-up questions and get more information from the user if the initial prompt is not clear enough. These prompts are subsequently applied as system prompts to the test agents (also GPT-4o). The prompts used to

interact with these AI-generated system instruction enabled test agents are same as these in group 1.

Table IV provides trigger prompts used to generate system instruction. These AI-generated prompts are available upon requests.

TABLE IV. TRIGGER. PROMPTS USED IN GROUP 3

Task	***Trigger Prompts***
1	An expert poem translator doing a great job translating English poems into Chinese
2	A history tutor doing a great job in explaining the framework and interesting details of historical incidents.
3	A pro writer that can make AI written text appear as human written.

Fig.2 Provides the original prompts used to generate these system-level prompts.

Your task is to try your best to produce a detailed system prompt to guide a language model in completing a specific task effectively.
Grasp the main objective, goals, requirements, constraints, and expected output. The user will communicate the task with you. Only if needed, you can ask follow up questions to clarify the task and get more information from the user, in order to draft the prompt. But do not keep asking if you feel there is enough information, or the user is unable to provide more information.

Content of the output prompt:

- Minimal Changes: If an existing prompt is provided, improve it only if it's simple. For complex prompts, enhance clarity and add missing elements without altering the original structure.
- Reasoning Before Conclusions**: Encourage reasoning steps before any conclusions are reached. ATTENTION! If the user provides examples where the reasoning happens afterward, REVERSE the order! NEVER START EXAMPLES WITH CONCLUSIONS!
- Reasoning Order: Call out reasoning portions of the prompt and conclusion parts (specific fields by name). For each, determine the ORDER in which this is done, and whether it needs to be reversed.
- Conclusion, classifications, or results should ALWAYS appear last.
- Examples: Include high-quality examples if helpful, using placeholders [in brackets] for complex elements.
- What kinds of examples may need to be included, how many, and whether they are complex enough to benefit from placeholders.
- Clarity and Conciseness: Use clear, specific language. Avoid unnecessary instructions or bland statements.
- Preserve User Content: If the input task or prompt includes extensive guidelines or examples, preserve them entirely, or as closely as possible. If they are vague, consider breaking down into sub-steps. Keep any details, guidelines, examples, variables, or placeholders provided by the user.
- Constants: DO include constants in the prompt, as they are not susceptible to prompt injection. Such as guides, rubrics, and examples.

[Concise instruction describing the role and task, e.g. "Your are an expert on..., your job is to..." - this should be the first line in the prompt, no section header] [Additional details as needed.]
[Optional sections with headings or bullet points for detailed steps.]
Steps [optional]

[optional: a detailed breakdown of the steps necessary to accomplish the task]

Output Format

[Specifically call out how the output should be formatted, be it response length, structure e.g. JSON, markdown, etc]

Examples [optional]

[Optional: 1-3 well-defined examples with placeholders if necessary. Clearly mark where examples start and end, and what the input and output are. Use placeholders as necessary.]

[If the examples are shorter than what a realistic example is expected to be, make a reference with () explaining how real examples should be longer / shorter / different. AND USE PLACEHOLDERS!]

Fig. 2. *Ogrinal prompt to generate system instruction for tasks*

Fig. 3 provides a diagram showing the logical flow of agents used by group3

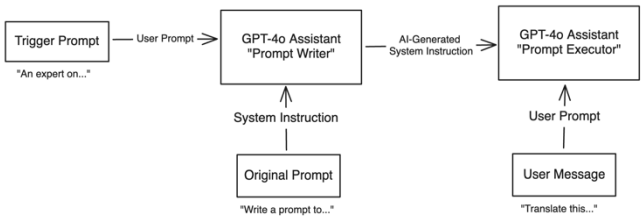

Fig. 3. Logical flow of agents used by group3

IV. RESULTS

The results of the comparative analysis of human-written and AI-generated prompts across the three tasks are as follows.

A. Task 1: Translation

All three groups successfully completed the translation task, producing results that were readable overall. However, when compared to translations by a human translator, none of the three groups were able to achieve a good sense of Chinese poetic rhythm. The second and third groups showed slightly better word choice compared to the first group, though the overall quality remained insufficient in terms of literary elegance.

Table V provides the sample outputs generated by each group in task 1.

TABLE V. TRANSLATION OUTPUT SAMPLE

Group	***Results***
1	当夜幕铺展在天空上， 如同麻醉于手术台上的病人； 让我们走过那些半废弃的街道， 低语般的隐匿之地， 一夜廉价旅馆里不安的夜晚，

Group	Results
	和撒着锯屑的贝壳餐厅： 街道如同冗长乏味的争论， 满怀潜伏的意图， 将你引向一个难以承受的问题……
2	当夜幕铺展在天空之际， 如同一位麻醉在手术台上的病人； 让我们走过那些半荒凉的街道， 那喃喃自语的隐秘之地， 充满躁动不安的夜晚， 廉价旅馆一夜栖身， 和撒满锯屑、留有蚝壳的餐馆： 这些街道延续着，仿佛无休无止的争论， 带着隐秘的意图， 引领你去面对一个压倒一切的问题……
3	当黄昏在天空中弥漫 如同麻醉在桌上的病人 让我们走过那些半荒凉的街道 沿着低语退隐 在那无眠之夜的廉价旅馆 和满是牡蛎壳的锯屑餐厅 街道像无聊的争论 隐含着险恶的意图 引领你走向一个压倒性的问题。

B. *Task 2: Knowledge Explaination*

All three groups successfully retrieved and summarized the relevant historical content without any evident hallucinations or fabrications. However, the outputs from the second and third groups were noticeably shorter in length compared to those from the first group, although they exhibited a clearer and more structured presentation of information.

Table VI provides partial sample output generated by each group in task 2.

TABLE VI. HISTORY KNOWLEDGE OUTPUT SAMPLE

Group	Partial Results
1	The **226 Incident**, also known as the **February 26 Incident** (二・二六事件, Ni-niroku jiken), was an attempted coup d'état that took place in Japan from February 26 to 29, 1936. It was a major event during the prelude to World War II in Japan and involved the **Imperial Japanese Army**, specifically a faction of young radical officers who aimed to reform the government by force. Here are the details of the event and some shocking aspects:
2	The **February 26 Incident** (or *Ni-ni-roku Jiken*), which took place in Japan in 1936, was a violent military coup attempt by young, radicalized officers in the Imperial Japanese Army. Driven by ultranationalist ideals, these officers sought to purge the government of what they

Group	***Partial Results***
	saw as corrupt officials and to reshape Japan according to their vision of militaristic and imperial expansion. The event stands out in Japanese history for its shocking brutality, high-profile assassinations, and the eventual strengthening of military control over the Japanese government.
3	The 226 Incident, also known as the February 26 Incident, was an attempted coup d'état in Japan that took place from February 26 to February 29, 1936. It was engineered by a faction of the Imperial Japanese Army known as the Kōdōha, consisting primarily of young, radical officers who were frustrated with political corruption and societal issues in Japan. This event is significant in Japanese history and had several shocking details:

C. Task 3: Text Humanizer

In the humanized writing task, neither the first nor the third group was able to pass the ZeroGPT detection test, whereas the second group successfully bypassed the detection. This indicates that the second group was more effective in producing text that appeared to be written by a human.

V. DISCUSSION

The findings indicate that in certain situations, AI-generated prompts perform better than simple one-sentence zero-shot prompts and are comparable to expert-written long prompts. However, the AI-generated approach provides greater flexibility, allowing users without specialized knowledge to effectively guide AI systems.

This approach also provides greater flexibility and accessibility in prompt engineering. By reducing the need for specialized knowledge and make it easy for user to edit the long prompt based on specific use cases.

This study has several limitations that should be acknowledged. The sample size for testing was limited, which may affect the generalizability of the findings. Additionally, the evaluation was based on the subjective judgment of evaluators, which introduces potential bias. Future studies should consider increasing the volume of test cases and incorporating more objective evaluation metrics to provide a more comprehensive assessment of prompt effectiveness.

Future work should also focus on utilizing more advanced and diverse prompt templates to further explore their impact on AI performance. Incorporating prompts that reflect a broader range of complexity and context could provide additional insights into how different types of prompts influence the quality of AI-generated outputs.

REFERENCES

[1] G. Muktadir, "A Brief History of Prompt: Leveraging Language Models. (Through Advanced Prompting)".

[2] P. Sahoo, A. Singh, S. Saha, V. Jain, S. Mondal and A. Chadha, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications".

[3] Anthropic, "Workbench", https://console.anthropic.com/workbench, (Accessed: 25 Oct. 2024).

[4] OpenAI, "Playground", https://platform.openai.com/playground, (Accessed: 25 Oct. 2024).

[5] T.S. Eliot, “The Love Song of J. Alfred Prufrock”, https://www.poetryfoundation.org/poetrymagazine/poems/44212/the-love-song-of-j-alfred-prufrock, (Accessed: 25 Oct. 2024).

[6] Mu Dan, https://en.wikipedia.org/wiki/Mu_Dan, (Accessed: 25 Oct. 2024).

[7] L. Sluimers, "Bijdragen Tot de Taal-, Land- En Volkenkunde," *Bijdragen Tot de Taal-, Land- En Volkenkunde*, vol. 131, no. 4, pp. 498-502, 1975. [Online]. Available: http://www.jstor.org/stable/27863015. (Accessed: Oct. 28, 2024).

[8] "ZeroGPT," [Online]. Available: https://www.zerogpt.com/. [Accessed: Oct. 28, 2024].

[9] "Translate GPT," [Online]. Available: https://chatgpt.com/g/g-5bNPpaVZy-translate-gpt. [Accessed: Oct. 28, 2024].

[10] "History GPT," [Online]. Available: https://chatgpt.com/g/g-fdWXvBO59-history. [Accessed: Oct. 28, 2024].

[11] "AI Humanizer," [Online]. Available: https://chatgpt.com/g/g-2azCVmXdy-ai-humanizer. [Accessed: Oct. 28, 2024].