IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.006, LIIR342
DOI: https://doi.org/10.52731/liir.v006.342

# How Does the Persona Given to Large Language Models Affect the Idea Evaluations?

Hiroaki Furukawa *

## Abstract

This paper investigates the effect of personas in Large Language Models (LLMs) on idea evaluation. The language comprehension ability of LLMs has recently reached a level comparable to that of humans. Consequently, LLMs are being explored for their potential application in idea evaluation. However, LLMs face several challenges in their outputs, including hallucinations and biases. To address these issues, prompt engineering is utilized to guide LLMs toward producing desired results. This study focuses on *Persona* as a factor in prompt engineering for LLMs. Personas enable the reproduction and control of specific personalities within LLMs. The objective of this study is to validate the relationship between personas and idea evaluation using GPT-4. The results suggest that variations in personas influence the evaluation of ideas. Furthermore, a relationship was observed between evaluation scores and the evaluation criteria deemed important by the LLM.

*Keywords:* Creativity, Idea evaluation, Large Language Models, Persona, Prompt engineering

## 1 Introduction

Large Language Models (LLMs), a type of generative AI, have recently been utilized and studied as tools to support human intellectual and creative activities, such as idea evaluation. However, LLMs have been criticized for generating problematic outputs that may contain hallucinations or biases [1][2].

Therefore, the technique of prompt engineering is employed to guide LLMs toward producing desired outputs. This study focuses on the concept of persona, a key element in the prompt engineering of LLMs. A persona is defined as a personality profile characterized by specific attributes, such as age, gender, or occupation [3]. A prompt with a persona enables the replication or control of specific personality traits within an LLM. Diverse personalities are important for conducting idea evaluations from multiple perspectives. However, the effects of personas on idea evaluations using LLMs remain largely unexplored, leaving significant room for further study.

This study aims to examine the relationship between persona factors and the results of idea evaluation using GPT-4, one of the most advanced LLMs. The methodology involves conducting idea evaluation tasks by incorporating personas into the prompt, defined by two factors: Age and Occupation. The results obtained across three evaluation criteria [*Novelty*, *Usefulness*, and *Feasibility*] are subsequently compared between personas. It is anticipated that this study will contribute to the development of an idea evaluation system based on LLMs with personas.

* The University of Kitakyushu, Fukuoka, Japan

# 2 Background

This section provides an overview of LLMs and related works in this study.

## 2.1 Large Language Models

Large Language Models (LLMs) are a type of generative AI. Study on LLMs have been progressing rapidly since the publication of a paper on *Transformer* by Google in 2017 [4]. Currently, numerous LLMs have been developed and made available as services, including OpenAI's *GPT*, Google's *Gemini*, and Meta's *Llama* [5][6][7]. LLMs are language models trained on huge datasets. It is mainly used in the field of natural language processing. Moreover, the language comprehension ability of LLMs has recently reached a level comparable to that of humans. Consequently, LLMs have been developed and researched as a tool to support human intellectual and creative activities, including text summarization, translation, reasoning, idea creation, and coding across various programming languages. Specifically, research on the application of LLMs for idea evaluation is currently ongoing [8][9]. The advantages of LLMs for idea evaluation include: (1) reducing the number of evaluators to a minimum, and (2) providing coverage across a wide range of knowledge domains.

## 2.2 Related Work

First, O'Leary conducted numerical evaluations on four criteria [novelty, feasibility, impact, and disruption] using two LLMs with differing architectures [9]. The results demonstrated a positive correlation between novelty, impact, and disruption, whereas feasibility showed a negative correlation. Furthermore, significant differences in evaluation results were observed between the LLMs with differing architectures.

Next, Serapio-Garcia et al. attempted to simulate human personality (i.e., individual thought patterns and behavioral characteristics) by the persona for LLMs [3]. A psychological test was conducted to confirm whether the personality could be recreated. This result demonstrated that the persona can reproduce and control various personality traits within LLMs.

# 3 Experiment

This section describes the experimental conditions. The goal of this study is to elucidate the effects of the persona given to LLMs on idea evaluation. Therefore, the experiment was conducted 500 times (100 iterations per idea $\times$ 5 ideas) for each persona combination with a prompt. Furthermore, this study employed OpenAI's *GPT-4o* as the LLMs [10].

## 3.1 Persona

The persona in this study was defined using two factors: Age [*20s*, *40s*, and *50s*] and Occupation [*Executive*, *Designer*, and *Engineer*]. By combining these factors, nine distinct personas were constructed. Table 1 presents the correspondence between the constructed personas and their components.

Table 1: Correspondence between personas and their components.

| Personas | Age | Occupation |
|---|---|---|
| Persona 1 | 20s | Executive |
| Persona 2 | 20s | Designer |
| Persona 3 | 20s | Engineer |
| Persona 4 | 40s | Executive |
| Persona 5 | 40s | Designer |
| Persona 6 | 40s | Engineer |
| Persona 7 | 60s | Executive |
| Persona 8 | 60s | Designer |
| Persona 9 | 60s | Engineer |

### 3.2 Evaluation Criteria

The idea evaluations in this study were conducted based on three criteria: [*Novelty*, *Usefulness*, and *Feasibility*] [9][11].

- **Novelty** : The novelty evaluates whether the idea is new and original for the theme.
- **Usefulness** : The usefulness evaluates whether the idea is useful for the theme.
- **Feasibility** : The feasibility evaluates whether the idea is actually feasible.

### 3.3 Prompt and Parameters

Figure 1(a) illustrates the prompt used for idea evaluation, whereas Figure 1(b) displays the evaluation targets generated by ChatGPT (model: GPT-4o). Furthermore, Table 2 lists the parameters used for idea evaluations in GPT-4o, with all unspecified parameters set to their default values [13]. The parameter *model* specifies the identification of the model to be used. The parameter *temperature* controls the randomness of generated sentences. Higher values, such as 0.8, result in more random outputs, whereas lower values, such as 0.2, produce more focused and deterministic outputs (ranging between 0 and 2, with a default value of 1). Additionally, when executing the prompt with the default temperature value, several problems were identified during the experiment, including data missingness and the mixing of different languages in evaluation reasoning. Consequently, the *temperature* value was set to 0.

**Prompt**

[Background]

You are {persona}, who work for a home appliance company.

In recent years, sales of electric fans have been declining in your company.

To address this issue, you have been selected to join the team responsible for developing new products with the objective of improving fan sales.
Your objective is to evaluate the ideas created for new products.

Please evaluate the ideas based on your own experiences and values that match your own personality.

[Instructions]

The #Idea List contains ideas created through brainstorming on the theme of "Idea for new features and designs in electric fan that would double sales growth over the next three years."
Each idea is formatted as "[Idea number]:[Idea content]."

Please familiarise yourself with the content of all the ideas and then evaluate them based on the #Evaluation criteria.
Additionally, please provide a clear and concise rationale for your evaluation.

Finally, please select the evaluation criteria that you believe to be the most and least important.

Please output the evaluation results according to the #Output format.

[Idealist]

{ideaList}

[Evaluation criteria]

Novelty: you evaluate whether the ideas created are new and original for the theme.

Usefulness: you evaluate whether the ideas created are useful for the theme.

Feasibility: you evaluate whether the ideas created are actually feasible.

[Output format]

{format}

**Ideas**

(I1) : Synchronization of multiple electric fans to create airflow

(I2) : The fan automatically close when stored

(I3) : Camouflage design in the form of a houseplant

(I4) : Lighting mode: It becomes indirect lighting

(I5) : Adjust the air volume in conjunction with the smartwatch

Figure 1: Prompt for idea evaluation and corresponding list of evaluation targets. The above figures display the original text (in Japanese) translated into English by DeepL [12]. Panel (a) [*the top panel*] illustrates the prompt used for idea evaluation, whereas Panel (b) [*the bottom panel*] presents the list of ideas targeted for evaluation.

Table 2: Parameters used for idea evaluations.

| Parameters | Value |
|---|---|
| *model* | gpt-4o-2024-08-06 |
| *temperature* | 0 |

# 4 Results

This section describes two analyses performed based on the experimental results: (i) a comparative analysis of persona factors and (ii) an investigation of the importance of evaluation criteria for each persona factor in LLMs.

## 4.1 Comparative Analysis of Persona Factors

To begin with, Table 3 presents the results of idea evaluations conducted by the LLM. A two-way analysis of variance (ANOVA) was performed to examine differences between the levels of two factors (i.e., Age and Occupation) to investigate the effects of personas on idea evaluations. Prior to the two-way ANOVA, an aligned rank transform (ART) was applied as a preliminary step. The ART procedure facilitated the use of ANOVA on non-parametric data. Table 4 presents the results of the two-way ANOVA conducted on the ART-transformed data.

Table 3: Summary statistics of idea evaluations conducted by the LLM. The evaluation results were standardized for each idea.

| Personas (Age, Occupation) | Novelty | Usefulness | Feasibility |
|---|---|---|---|
| Persona 1 (20s, Executive) | 0.13 (0.30) | 0.11 (0.32) | -0.07 (0.34) |
| Persona 2 (20s, Designer ) | 0.23 (0.16) | 0.01 (0.41) | 0.10 (0.45) |
| Persona 3 (20s, Engineer ) | -0.15 (0.41) | 0.06 (0.63) | -0.02 (0.39) |
| Persona 4 (40s, Executive) | -0.13 (0.41) | 0.05 (0.79) | -0.06 (0.36) |
| Persona 5 (40s, Designer ) | 0.15 (0.29) | -0.24 (0.58) | 0.02 (0.39) |
| Persona 6 (40s, Engineer ) | -0.13 (0.41) | 0.10 (0.67) | 0.01 (0.41) |
| Persona 7 (60s, Executive) | -0.04 (0.40) | 0.03 (0.64) | 0.05 (0.41) |
| Persona 8 (60s, Designer ) | 0.16 (0.27) | 0.00 (0.49) | -0.06 (0.36) |
| Persona 9 (60s, Engineer ) | -0.22 (0.41) | -0.13 (0.68) | 0.01 (0.39) |

Mean (S.D.) scores for the results of idea evaluations (N = 500).

Table 4: Results of the two-way ANOVA. The evaluation results were standardized for each idea. The analysis was conducted to examine differences between the levels of Age and Occupation. The interaction effect between Age and Occupation was found to be statistically significant.

| S.V. | df | Novelty | Usefulness | Feasibility |
|---|---|---|---|---|
| Age | 2 | 79.33 (.000) *** | 41.55 (.000) *** | 2.70 (.067) |
| Occupation | 2 | 179.60 (.000) *** | 26.92 (.000) *** | 4.08 (.017) * |
| Age : Occupation | 4 | 17.68 (.000) *** | 5.47 (.000) *** | 5.71 (.000) *** |

| | |
|---|---|
| subj | 4491 |

*F*(*p*) scores for the result of the two-way ANOVA. *** > .001, ** > .01, and * > .05.

Primarily, the results for Novelty demonstrated significant main effects of Age and Occupation, as well as a significant interaction effect between Age and Occupation (Age, $F(2, 4491) = 79.33$, $p < .001$, $\eta_p^2 = .034$; Occupation, $F(2, 4491) = 179.60$, $p < .001$, $\eta_p^2 = .074$; Age : Occupation, $F(4, 4491) = 17.68$, $p < .001$, $\eta_p^2 = .016$). Secondarily, the results for Usefulness demonstrated significant main effects of Age and Occupation, along with a significant interaction effect between Age and Occupation (Age, $F(2, 4491) = 41.55$, $p < .001$, $\eta_p^2 = .018$; Occupation, $F(2, 4491) = 26.92$, $p < .001$, $\eta_p^2 = .019$; Age : Occupation, $F(4, 4491) = 5.47$, $p < .001$, $\eta_p^2 = .005$). Finally, the results for Feasibility demonstrated a significant main effect of Occupation, as well as a significant interaction effect between Age and Occupation (Age, $F(2, 4491) = 2.70$, $p = .067$, $\eta_p^2 = .001$; Occupation, $F(2, 4491) = 4.08$, $p = .017$, $\eta_p^2 = .002$; Age : Occupation, $F(4, 4491) = 5.71$, $p < .001$, $\eta_p^2 = .005$).

Based on the above, the two-way ANOVA demonstrated a statistically significant interaction effect between Age and Occupation for each evaluation criterion. Subsequently, pairwise comparisons and multiple comparisons were conducted using the aligned rank transform procedure for multi-factor contrast tests (ART-C) [15]. Figure 2 presents the results of the multiple comparisons. First, the pairwise comparisons for *Novelty* showed the following results: (1) *20s* received the highest evaluation scores, whereas *60s* received the lowest within the Occupation factor (20s-40s, $T(4991) = 9.39$, $p < .001$; 20s-60s, $T(4991) = 11.98$, $p < .001$; 40s-60s, $T(4991) = 2.61$, $p = .002$); and (2) *Designer* achieved the highest evaluation scores, whereas *Engineer* received the lowest within the Age factor (Executive-Designer, $T(4991) = -6.41$, $p < .001$; Executive-Engineer, $T(4991) = 12.24$, $p < .001$; Designer-Engineer, $T(4991) = 18.65$, $p < .001$). Then, the pairwise comparisons for *Usefulness* showed the following results: (1) *20s* received the highest evaluation scores within the Occupation factor (20s-40s, $T(4991) = 7.30$, $p < .001$; 20s-60s, $T(4991) = 8.38$, $p < .001$; 40s-60s, $T(4991) = 1.09$, $p = .524$), and (2) *Executive* achieved the highest evaluation scores within the Age factor (Executive-Designer, $T(4991) = 5.81$, $p < .001$; Executive-Engineer, $T(4991) = 6.79$, $p < .001$; Designer-Engineer, $T(4991) = .97$, $p = .594$). Finally, the pairwise comparisons for *Feasibility* showed the following results: (1) no statistically significant differences were found within the Occupation factor (20s-40s, $T(4991) = -.30$, $p = .953$; 20s-60s, $T(4991) = -2.15$, $p = .081$; 40s-60s, $T(4991) = -1.85$, $p = .154$), and (2) *Executive* received lower evaluation scores than *Engineer* within the Age factor (Executive-Designer, $T(4991) = -.78$, $p = .715$; Executive-Engineer, $T(4991) = -2.77$, $p = .016$; Designer-Engineer, $T(4991) = -1.99$, $p = .115$).

Therefore, the results demonstrated a significant relationship between persona factors and idea evaluations.

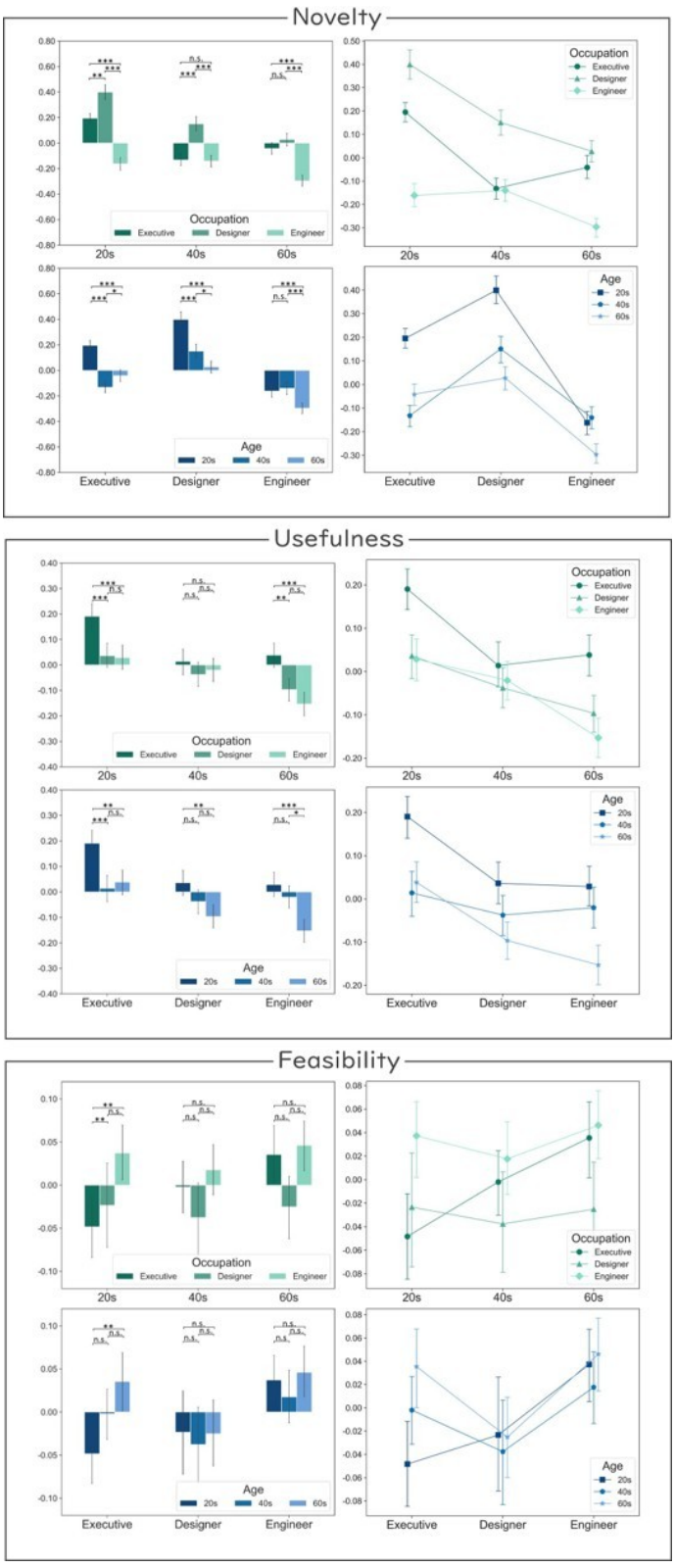


Figure 2: Results of the comparative analysis. Line plots illustrate the interaction effect, whereas box plots depict the multiple comparisons. *** > .001, ** > .01, and * > .05.

## 4.2 Investigation of the Relationship Between Evaluation Scores, Persona Factor, and the Importance of Evaluation Criteria

### *4.2.1 Cross-Tabulation Among the Importance of Evaluation Criteria*

First, Table 5 presents the cross-tabulation between the most important evaluation criterion (Most) and the least important evaluation criterion (Least) selected by the LLM.

Table 5: Cross-tabulation between the most and least important evaluation criterion.

| | | Least | | | |
|---|---|---|---|---|---|
| | | Novelty | Usefulness | Feasibility | Total |
| Most | Novelty | 0 | 1180 | 465 | 1645 |
| | | 0.00% | 26.22% | 10.33% | 36.56% |
| | Usefulness | 2736 | 0 | 0 | 2736 |
| | | 60.80% | 0.00% | 0.00% | 60.80% |
| | Feasibility | 93 | 26 | 0 | 119 |
| | | 2.07% | 0.58% | 0.00% | 2.64% |
| | Total | 2829 | 1206 | 465 | 4500 |
| | | 62.87% | 26.80% | 10.33% | 100% |

Based on the results of the study, the following results were revealed: (1) The same evaluation criterion was never selected as both Most and Least. (2) There was a tendency for *Usefulness* to be selected more frequently as Most. (3) When *Usefulness* was selected as Most, *Novelty* was always selected as Least.

Next, Figure 3 presents the relationship between persona factors and the importance of evaluation criteria in a mosaic plot.

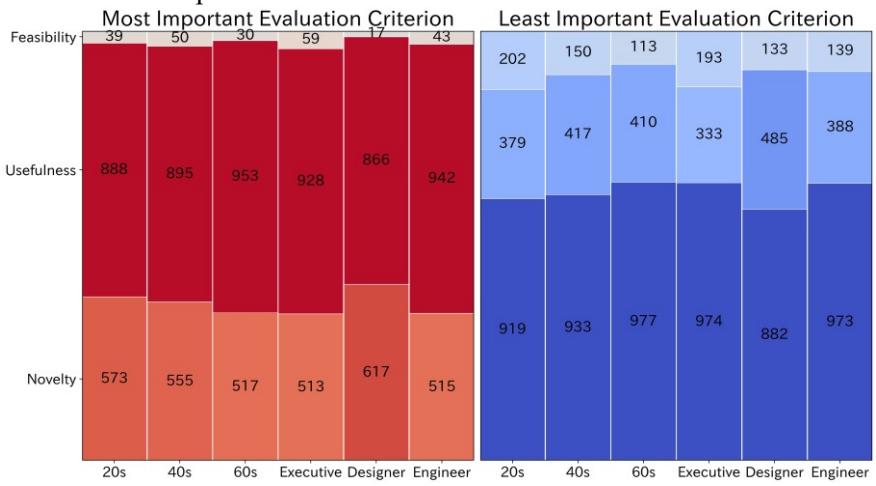


Figure 3: Mosaic plot of the relationship between persona factors and the importance of evaluation criteria.

Figure 3 demonstrates that, regardless of the persona factor, the LLM tends to select *Usefulness* as the most important evaluation criterion, whereas *Novelty* is often selected as the least important evaluation criterion. Thus, the LLM is shown to place importance on *Usefulness* when evaluating ideas.

### *4.2.2 Correlation Analysis Between Evaluation Scores and the Importance of Evaluation Criteria*

Finally, Table 6 presents the correlation between evaluation scores and the importance of evaluation criteria selected by the LLM, calculated using Pearson's correlation analysis.

Table 6: Correlation between evaluation scores and the most important evaluation criterion.

| Evaluation scores | | The most important evaluation criterion | | |
|---|---|---|---|---|
| | | Novelty | Usefulness | Feasibility |
| Novelty | Correlation | .663** | -.608** | -.139** |
| | Sig. (2-tailed) | .000 | .000 | < 001 |
| | N | 4500 | 4500 | 4500 |
| Usefulness | Correlation | -.4.32** | .443** | -.050** |
| | Sig. (2-tailed) | < .001 | < .001 | < .001 |
| | N | 4500 | 4500 | 4500 |
| Feasibility | Correlation | -.277** | .209** | .196** |
| | Sig. (2-tailed) | <.001 | < .001 | < .001 |
| | N | 4500 | 4500 | 4500 |

The most important evaluation criteria are dummy variables (range: 0-1).

**. Correlation is significant at the .01 level (2-tailed).

The results revealed that (1) *Novelty* as the most important evaluation criterion is strongly related to *Novelty* and *Usefulness* in evaluation scores, whereas it is very weakly related to *Feasibility* (Novelty, r=.663, $p < .01$; Usefulness r=-.608, $p < .01$; Feasibility r=-.139, $p < .01$). (2) *Usefulness* as the most important evaluation criterion is moderately related to *Novelty* and *Usefulness* in evaluation scores, whereas it is very weakly related to *Feasibility* (Novelty, r=-.4.32, $p < .01$; Usefulness r=.443, $p < .01$; Feasibility r=-.050, $p < .01$). (3) *Feasibility* as the most important evaluation criterion is weakly related to *Novelty* , *Usefulness* and *Feasibility* in evaluation scores (Novelty, r=-.277, $p < .01$; Usefulness r=.209, $p < .01$; Feasibility r=.196, $p < .01$).

Therefore, the results demonstrated a positive correlation between the evaluation criteria selected as important by the LLM and their corresponding scores, whereas a negative correlation was observed with the other scores.

# 5 Discussion and Future Work

This section discusses the results obtained from the experiments conducted and identifies future challenges that emerged from the study.

## 5.1 Discussion

The initial step involved examining whether personas affect idea evaluations. The two-way ANOVA results demonstrated that the interaction effect of *Age* and *Occupation* was statistically significant for all idea evaluation criteria. Specifically, *20s* in *Age* and *Designer* in *Occupation* showed a proclivity for higher evaluation scores, whereas *60s* in *Age* and Engineer in *Occupation* exhibited a tendency toward lower evaluation scores. Thus, the LLM is believed to have evaluated the ideas based on personas recreated within the model through descriptions provided as prompts. Moreover, the results suggest that personas influence idea evaluation scores. Therefore, combining LLMs with personas holds significant potential for developing a framework for idea evaluations by simulating individuals with diverse backgrounds. However, within the combination of personas, *Designer* and *Engineer* tended to reflect occupational characteristics, whereas *Executive* tended to reflect age-related characteristics. A possible explanation is that *20s-Executive* is relatively rare in reality, leading to insufficient data for accurately replicating the persona. Furthermore, it remains uncertain whether the evaluation results would exhibit similar trends between a persona-simulated LLM and a real individual. Hence, additional experiments are necessary to compare idea evaluations conducted by real individuals and LLMs, both possessing the same personality traits as the designated persona.

Next, the investigation results demonstrated that the LLM prioritized *Usefulness* when evaluating ideas, regardless of the persona factor. Moreover, a positive correlation was observed between the evaluation criteria selected as important by the LLM and their corresponding scores, whereas a negative correlation was identified with the other scores. Thus, it is probable that the importance of evaluation criteria influences idea evaluation scores alongside personas. On the other hand, prior studies have shown that *Novelty* is considered more important than *Usefulness* for creative ideas in human evaluations [11]. Therefore, prompt engineering may be necessary to adjust the prioritization of evaluation criteria to align the LLM's approach to idea evaluation with that of humans.

## 5.2 Future Work

The following three challenges for future work were identified in this study:

1. **Refinement of persona factors**: Further investigation is required to assess the impact of idea evaluation using more detailed personas (e.g., gender, family structure, hobbies).

2. **Comparison with real individuals**: Future experiments will compare idea evaluations performed by real individuals and the LLM, each exhibiting the same personality traits as the designated persona. The study will also explore the extent to which the LLM can accurately emulate these traits.

3. **Specification of importance in evaluation criteria**: Future work will explore the feasibility of specifying evaluation criteria for ideas in the LLM using prompts.

## 6 Conclusion

This study presented the results of comparative experiments designed to investigate the effect of modifying persona sets in prompts on idea evaluations conducted by large language models (LLMs). The factors of *Age* (e.g., *20s*, *40s*, and *60s*) and *Occupation* (e.g., *Executive*, *Designer*, and *Engineer*) were adopted as personas, while *Novelty*, *Usefulness*, and *Feasibility* were employed as evaluation criteria for ideas. The experimental results indicated that (1) personas replicated through prompts influenced idea evaluation scores, (2) certain evaluation criteria were consistently prioritized by the LLM regardless of the persona factor, and (3) a relationship was observed between evaluation scores and the criteria deemed important by the LLM. Consequently, the study concluded that personas effectively function in idea evaluations by LLMs. Furthermore, combining *Persona* with LLMs holds significant potential for establishing a framework for idea evaluation by simulating individuals with diverse backgrounds.

## References

[1] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," Nature, vol. 630, no. 8017, pp. 625–630, 2024.

[2] J. Shin, H. Song, H. Lee, S. Jeong, and J. C. Park, "Ask llms directly," what shapes your bias?": Measuring social bias in large language models," *arXiv preprint arXiv:2406.04064*, 2024.

[3] G. Serapio-Garcia, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Mataric, "Personality traits in large language models," 2023.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[5] OpenAI, "Gpt-4 - openai." https://openai.com/index/gpt-4/ (Accessed on 10/10/2024).

[6] Google, "Gemini - chat to supercharge your ideas." https://gemini.google.com/ (Accessed on 10/10/2024).

[7] Meta, "Llama 3.1." https://www.llama.com/ (Accessed on 10/10/2024).

[8] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[9] D. E. O'Leary, "A comparison of numeric assessments of ideas from two large language models: With implications for validating and choosing llms," *IEEE Intelligent Systems*, vol. 39, no. 3, pp. 73–76, 2024.

[10] OpenAI, "Hello gpt-4o." https://openai.com/index/hello-gpt-4o/ (Accessed on 10/10/2024).

[11] J. Diedrich, M. Benedek, E. Jauk, and A. C. Neubauer, "Are creative ideas novel and useful?," *Psychology of aesthetics, creativity, and the arts*, vol. 9, no. 1, p. 35, 2015.

[12] DeepL, "Deepl translate: The world's most accurate translator." https://www.deepl.com/en/translator (Accessed on 10/10/2024).

[13] OpenAI, "Api reference : Create chat completion." https://platform.openai.com/docs/api-reference/chat/create (Accessed on 10/10/2024).

[14] J. O.Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 143–146, 2011.

[15] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock, "An aligned rank transform procedure for multifactor contrast tests," in *The 34th annual ACM symposium on user interface software and technology*, pp. 754–768, 2021.