IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.004, LIIR168
DOI: https://doi.org/10.52731/liir.v004.168

Qualitative Evaluations of Ideas Created by Generative AI

Hiroaki Furukawa *

Abstract

This paper describes that the difference between GPT-3 and human in the qualitative evaluation of ideas. GPT-3 is expected to be used as 'Artificial General Intelligence' that requires no pre-training for a specific applications or purposes different from the conventional language model. This study aims to validate the usefulness of the ideas created by GPT-3. GPT-3 was validated using three pre-training approach; zero-shot, one-shot, and few-shot. The qualitative evaluation of ideas was conducted on three items; fluency, feasibility and originality. The comparative experiments were conducted on the evaluation results of GPT-3-created ideas and human-created ideas, as well as on the results of the in-context learning settings for tasks in GPT-3. The results suggested that human-created ideas were superior to GPT-3-created ideas in originality; moreover, few-shot is the highest of the approaches in originality.

Keywords: Creativity, Generative AI, GPT-3, Idea evaluation, Qualitative evaluation.

1 Introduction

GPT-3, a model for natural language processing in 'Artificial Intelligence (AI)', was published by OpenAI in 2020 [1]. Moreover, it has had a major impact on sentence generation in Generative AI. Conventional AI technologies have been considered difficult to use due to the need to customize learning contents and build highly specialized AI for each purpose or application. Whereas, GPT-3 is universally applicable without learning for individual purposes and applications. Therefore, GPT-3 is believed to have been realized as a 'Artificial General Intelligence (AGI)' [2]. GPT-3 has enabled AI to perform the intellectual creative activities represented by idea creation. However, GPT-3-generated text has a possibility to occur ethically problematic results or make no sense. [3]. Furthermore, the effect of the relationships between the quality of ideas and the in-context learning settings in GPT-3, as well as the relationships among evaluation items remain unclear.

This study aimed to validate the usefulness of the ideas created by GPT-3. The method is to perform the qualitative evaluations of ideas between the GPT3-created ideas and the human-created ideas; subsequently, a comparative experiment is conducted. Therefore, GPT-3 is expected to support creative activities in the future.

2 Background

This section presents an overview of GPT-3, as well as related work between GPT-3 and creativity.

* The University of Kitakyushu, Fukuoka, Japan

2.1 OpenAI

OpenAI is an AI research and deployment company [4]. Their mission is to ensure that artificial general intelligence benefits all of humanity. ChatGPT is the most famous AI product developed by OpenAI. The natural language processing model used by ChatGPT is 'Generative Pre-Trained Transformer (GPT)'. OpenAI technology is freely available to all. Therefore, they contribute to the advancement of AI.

2.2 GPT-3

'Generative Pre-Trained Transformer (GPT)' is autoregressive language model that uses deep learning to produce human-like text; moreover, GPT-3 is a third-generation released by OpenAI in 2020 [1][2][3]. GPT-3 is characterized by the very large training parameters included in the language model. The first iteration of GPT in 2018 used 110 million learning parameters. GPT-2 in 2019 used 1.5 billion of them, GPT-3 uses 175 billion parameters. Thus, the parameters of autoregressive language model have increased over the years. In addition, GPT-3 uses very large datasets (570 GB from 45TB of textdata) to pre-train language models. GPT-3 is enabled to automatically create sentences as a human with high accuracy prediction of the next word based on huge datasets. Note that, the latest model at the time of writing this paper is GPT-4.

2.3 Related work

Stevenson, Smal, Baas, Grasman, and van der Maas suggested that the creativity of GPT-3 was assessed on Guilford's Alternative Uses Test (AUT) [5]. They compared creativity between GPT-3 and human with five items; originality, usefulness, surprise of responses, flexibility and semantic distance between a response and the AUT object in question. These results concluded that human creativity is better than GPT-3 in originality, surprise, and semantic distance. Their study was to clarify the difference in the quality of ideas between GPT-3 and human; whereas, the relationships between the quality of ideas and the in-context learning settings in GPT-3, as well as the relationships among evaluation items are not clear.

3 Method

This study goal is to validate the usefulness of the ideas created by GPT-3. Therefore, the author focused on Pre-training approach, including the in-context learning settings in GPT-3. This study identifies the effect of setting differences on the quality of ideas.

3.1 Pre-training approach

First, the four pre-training approaches are presented in GPT-3 task execution [1].

- **Zero-Shot (0S)** is only given a natural language instruction describing the task. This approach provides maximum convenience, potential for robustness, and avoidance of 'spurious correlations' (e.g., "*As population increases, crime rates increase. As the population*

increases, the stores are increasing. Thus, more crime rate means more the stores." Is this relationship really correct?). Whereas, zero-shot is difficult to understand the form of the task without prior examples. In this case, the output may be completely different from the expected response.

- **One-Shot (1S)** is only one demonstration is allowed, in addition to a natural language description of the task. The reason is that this approach most closely matches the way some tasks are communicated to humans.
- **Few-Shot (FS)** is given a few demonstrations of the task at inference time as conditioning. It works by giving K examples of context and completion (K in the range of 10 to 100). The merits are a significant reduction in the need for task-specific data, as well as a reduced potential for learning an overly narrow distribution from a large dataset. Whereas, this approach is far inferior to fine-tuned models.
- **Fine-Tuning (FT)** is the most common approach in recent years. This approach requires thousands to hundreds of thousands of labeled examples as a supervised dataset specific to the desired task to update the weights of a pre-trained model. The merit is strong performance on many benchmarks. The most significant demerit is the need for a new large dataset for every task.

Fine-tuning demonstrates excellent performance in many benchmarks. However, preparing hundreds of thousands of labeled data for every task is not realistic. Therefore, three approaches were adopted in this study; 'Zero-Shot', 'One-Shot', and 'Few-Shot'.

3.2 Procedure and environment

Second, the experimental procedure is described idea creation using GPT-3. The experiments were conducted in the three settings for each pre-training approach using GPT-3; 'Zero-Shot (0S)', 'One-Shot (1S)', and 'Few-Shot (FS)'.

Table 1: **The parameters for creating sentence in GPT-3.** *The all parameters that not listed below are defaults*.

Parameters	Value
model	text-davinci-002
temperature	0.8
max_tokens	4000
presence_penalty	1.
frequency_penalty	1.

Table 1 displayed the parameters for creating ideas in GPT-3 [4]. ‘model’ indicates the ID of the model to use GPT-3. ‘temperature’ indicates the randomness of generated sentences. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic (Between 0 and 2, Defaults to 1). ‘max_tokens’ indicates the maximum number of tokens to generate in the chat completion. The total length of input tokens and generated tokens is limited by the model's context length. ‘presence_penalty’ indicates the positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics (Between -2.0 and 2.0, Defaults to 0). ‘frequency_penalty’ indicates the positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim (Between -2.0 and 2.0, Defaults to 0).

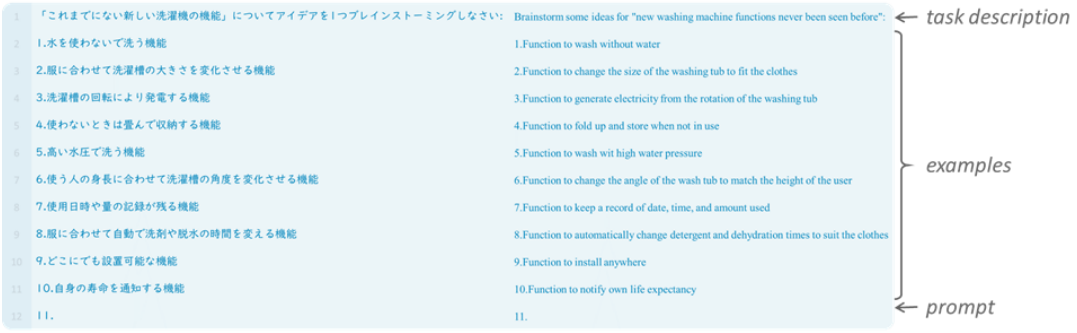


Figure 1: **Example of the GPT-3 task in this study.** *The original text was translated into English at DeepL* [6].

Figure 1 displayed the example of the GPT-3 task in this study. The panels above display the example of the GPT-3 task in this study. The left side displays original (written in Japanese); moreover, the right-side displays translated into English at DeepL. Row 1 was task description. Zero-shot used no examples, one-shot used only 1 example (Row 2), and few-shot used 10 examples (Rows 2 through 11). Row 12 was prompt.

After performing the above task one hundred times — one hundred ideas created — for each setting, the fifty ideas were randomly selected. The human-created ideas for comparison with GPT-3-created ideas were randomly selected from the previous experiment [7]. The reason for the random selection was to reduce the burden on the judges.

3.3 Measures

And finally, the three evaluation items are presented in ideas [8][20].

- **Fluency** evaluates whether the ideas are appropriate for the task.
- **Feasibility** evaluates whether the ideas are actually feasible for the task.
- **Originality** evaluates whether ideas are unique for the task that never seen before.

12 judges (Female: 9, Male: 3, Average age: 23.36, Nationality: Japanese) evaluated each idea on fluency, feasibility, and originality using 4-point scale (from 1 = "Strongly disagree" to 4 = "Strongly agree"). The fluency | feasibility | originality scores were the sum of each evaluation results. The Scores were calculated for each idea. The judges were blinded to whether or not the responses stemmed from humans or GPT-3. The evaluation was conducted using a web-based questionnaire.

4 Result

This section describes that three analyses were performed based on the results of the experiments; text mining, the comparison of ideas evaluation, and correlation analysis.

4.1 Overview: Examples of ideas

To begin with, Table 2 and Table 3 display the examples of ideas created in this study. Since all judges were Japanese and GPT-3 had been given tasks in Japanese, all ideas have been created in Japanese. Therefore, English translations were provided using DeepL [6].

Table 2: **Examples of ideas created by GPT-3.** *Three ideas are displayed for each setting.*

Setting	Original	Translation
Zero-Shot (0S)	タッチパネル操作	Touch panel
	自動で洗剤を入れる機能	Automatic detergent injection function
	洗濯時間を最短化する機能	Functions to minimize washing time
One-Shot (1S)	直接除菌する機能	Direct sterilization function
	電気消費量を抑える機能	Functions to reduce electricity consumption
	自動で衣服を洗う機能	Automatic clothes washing function
Few-Shot (FS)	洗い方を動画で教える機能	Video teaching function on how to wash
	洗剤を節約する機能	Detergent-saving features
	洗剤を選ぶ機能	Detergent selection function

Table 2 columns are explained: The ‘Setting’ column indicates the in-context learning settings for tasks in GPT-3. The ‘Original’ column indicates examples of ideas created by GPT-3. GPT-3 had been given tasks in Japanese, thus ideas were created in Japanese. The ‘Translation’ column indicates examples of ideas in the ‘Original’ column translated into English at DeepL.

Table 3: **Examples of ideas created by Humans.** *Three ideas are displayed from human-created ideas.*

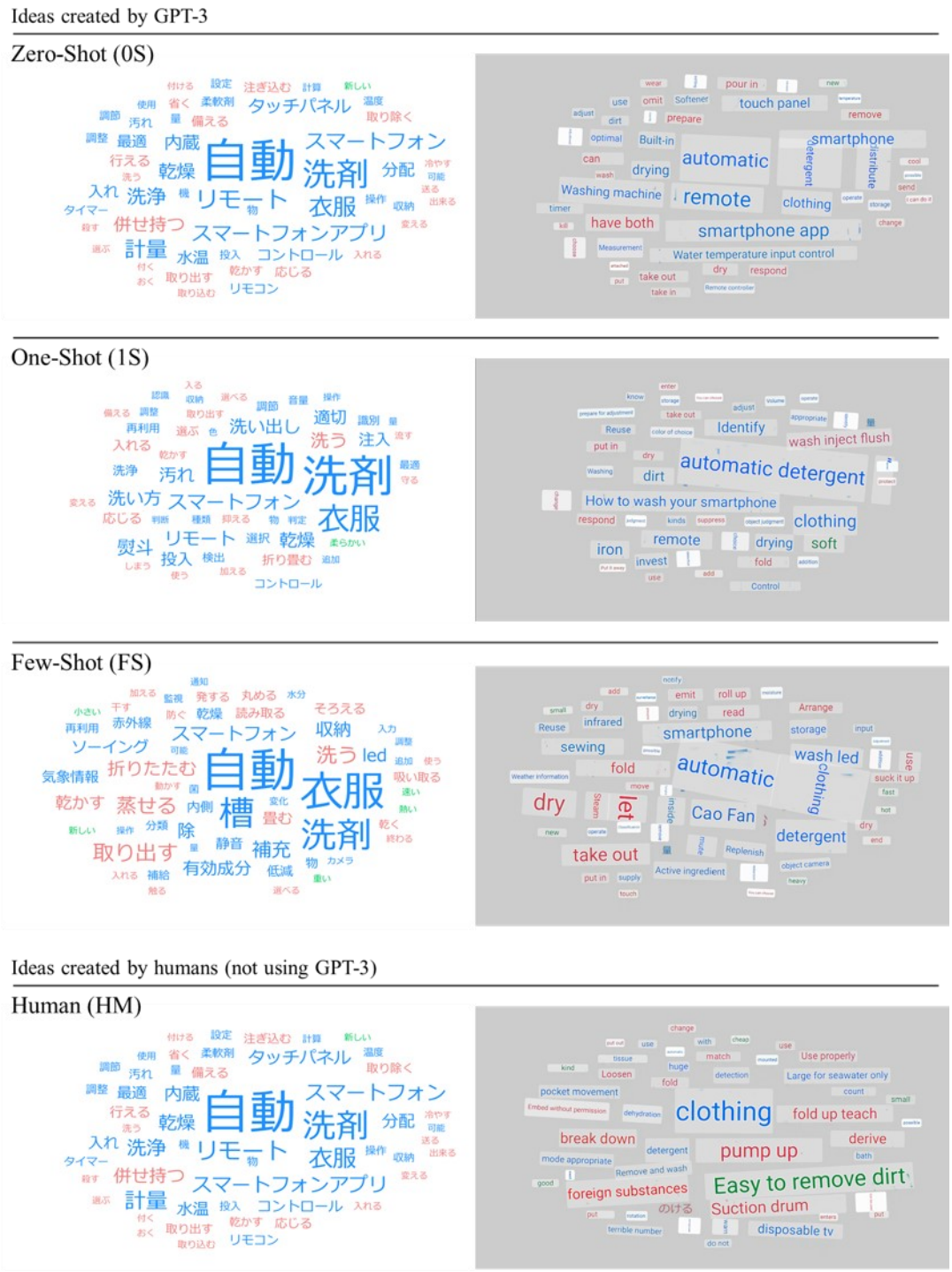
Original	Translation
音声入力機能	Voice input function
服に合わせて自動で洗剤や脱水の時間が変わる洗濯機	Washing machine that automatically changes detergent and dehydration time according to the clothes
海水を利用できる	Seawater available

Table 3 columns are explained: The ‘Original’ column indicates examples of human-created ideas, thus ideas were created in Japanese language. The ‘Translation’ column indicates examples of ideas in the ‘Original’ column that have been translated into English at DeepL.

4.2 Text mining

Next, text mining was conducted to analyze the words in the ideas. Figure 2 displayed the results of the analysis visualized as a word cloud. The panels display four word clouds of the ideas. The left-side displays the original ideas; moreover, the right-side displays the ideas translated using Google Lens. ‘Zero-Shot (0S)’ displays a word cloud of the ideas created by GPT-3 using zero-shot setting. ‘One-Shot (1S)’ displays a word cloud of the ideas created by GPT-3 using one-shot setting. ‘Few-Shot (FS)’ displays a word cloud of the ideas created by GPT-3 using few-shot setting. In addition, zero-, one-, and few-shot are the in-context learning for tasks in GPT-3. ‘Human (HM)’ displays a word cloud of the ideas created from brainstorming by humans. The panels displayed that some words were observed to have a high frequency of occurrence in common among 0S, 1S, and FS; moreover, HM has more words and parts of speech in the ideas than 0S, 1S and FS.

A ‘word cloud’ (also known as a ‘tag cloud’) is a method of displaying a list of words in a spatial layout by weighting the words in a sentence according to their frequency of occurrence and importance [10][11][12]. A word cloud provides a visual representation of the results of the text mining analysis [13][14]. The text mining tool in this study uses the ‘TF.IDF (Term Frequency times Inverse Document Frequency) method’ to express the importance of words [15]. The TF.IDF method is calculated as the product of two indices: TF (Term Frequency) is the frequency with which a word occurs in a given sentence, and IDF (Inverse Document Fre-



quency) is
the frequency with which a word occurs in all sentences.

Figure 2: **Comparison of word clouds.** *Word clouds created by a text mining tool from User Local, Inc. [9]. The word clouds in this figure are displayed based on scores using the TF.IDF method.*

As shown in Figure 2, word clouds are displayed based on the frequency of occurrence of a

word and the importance of the word as calculated by the TF.IDF method; moreover, characteristic words are labelled with a larger font size, otherwise uncharacteristic are labelled with a smaller font size. In addition, the color of the label displays the part of speech (e.g., 'nouns' are blue, 'verbs' are red, and 'adjectives' are green).

The analysis results suggest that some words were observed to have a high frequency of occurrence in common among 'Zero-Shot (0S)', 'One-Shot (1S)', and 'Few-Shot (FS)'; moreover, 'Human (HM)' had more words and parts of speech in the ideas than 0S, 1S and FS.

4.3 Comparison

Then, Table 4 displayed the results of the ideas evaluation, Figure 3, Figure 4, and Figure 5 also displayed the comparison of the ideas evaluation and variances. Note that, since the results of each evaluation were not expected to follow a normal distribution, the median was used as the central tendency.

Table 4: **The results of the ideas evaluation.** *A number in this table is given as a median (range = min - max).*

	Fluency	Feasibility	Originality
Zero-Shot (0S) (n=50)	33.5 (20-44)	43.0 (23-48)	25.5 (15-42)
One-Shot (1S) (n=50)	35.0 (18-44)	38.5 (21-48)	30.0 (14-44)
Few-Shot (FS) (n=50)	34.5 (24-44)	35.0 (17-48)	32.5 (17-45)
Human (HM) (n=50)	40.0 (28-46)	30.0 (15-45)	40.0 (21-46)

Table 4 demonstrated that 'Human (HM)' tends to have a narrower range and higher median than the other groups in 'Fluency' and 'Originality', whereas HM tends to have a wider range and lower median in 'Feasibility'.

In order to clarify the statistically significant differences among the groups in the results of the ideas evaluation, the Kruskal-Wallis test was used in this study (with a significance level of .05 to determine). The Kruskal-Wallis test detected the statistically significant differences among the four groups in 'Fluency' ($p < .001$), 'Feasibility' ($p < .001$) and 'Originality' ($p < .001$). Since statistically significant differences were found among the groups, the Steel-Dwass test was conducted as a multiple comparison test (with a significance level of .05 to determine). The Steel-Dwass test detected the statistically significant differences across the groups (Figure 3, Figure 4 and Figure 5).

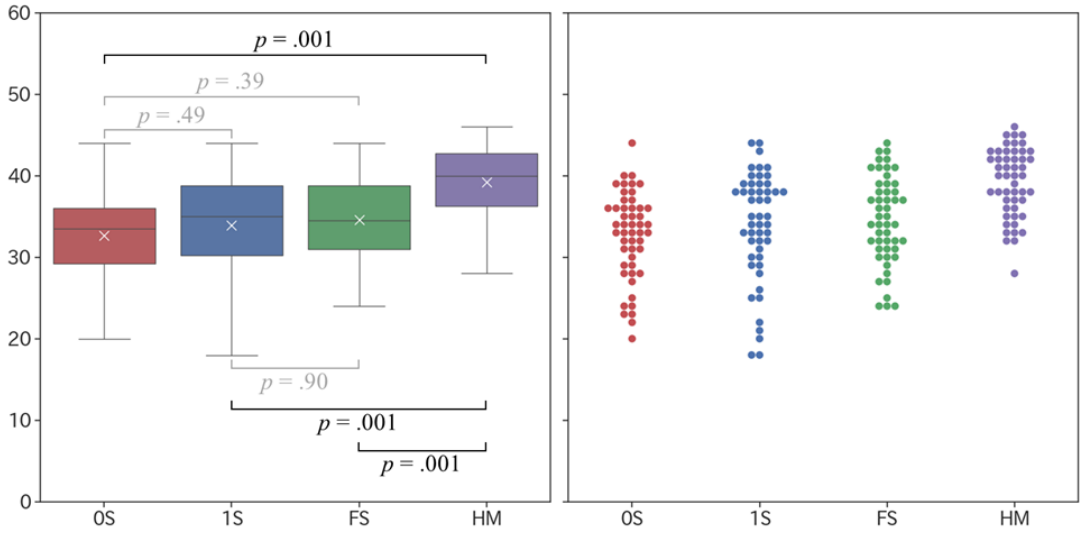


Figure 3: **Comparison of fluency.** *The left panel displays a boxplot of fluency in the ideas; moreover, the right panel displays a swarmplot.*

Primarily, Figure 3 displayed that the statistically significant differences were found between 'Human (HM)' and the other groups in fluency of the ideas. The results demonstrated that HM tends to have the highest of groups in fluency. Whereas, no significant differences were found among 'Zero-Shot (0S)', 'One-Shot (1S)', and 'Few-Shot (FS)', i.e., these groups used GPT-3.

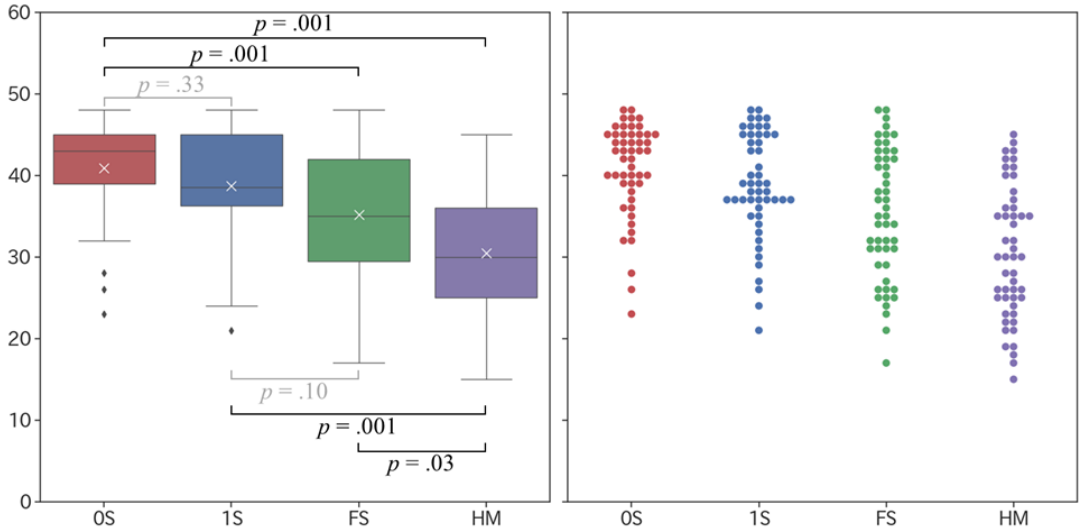


Figure 4: **Comparison of feasibility.** *The left panel displays a boxplot of feasibility in the ideas; moreover, the right panel displays a swarmplot.*

Secondary, Figure 4 displayed that the statistically significant differences were found between HM and the other groups, and between 0S and FS in feasibility of the ideas. The results demonstrated that HM tends to have the lowest of the groups in feasibility. Likewise, FS tends to have lower scores than 0S.

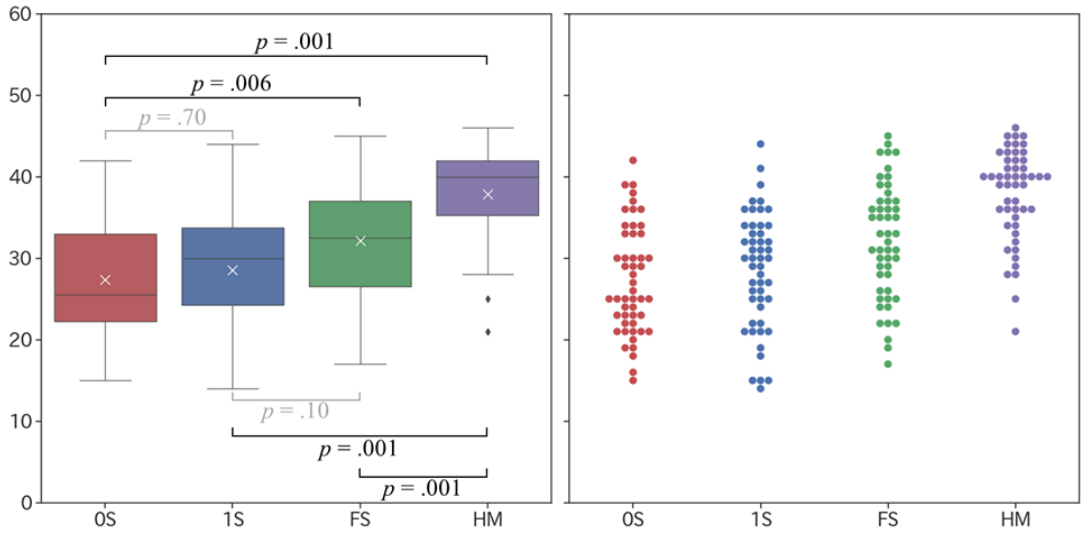


Figure 5: **Comparison of originality.** *The left panel displays a boxplot of originality of the ideas; moreover, the right panel displays the swarmplot.*

Tertiary, Figure 5 displayed that the statistically significant differences was found between HM and the other groups, and between 0S and FS in originality of the ideas. The results demonstrated that HM tends to have the highest of the groups in originality. Likewise, FS tends to have higher scores than 0S.

4.4 Correlation analysis

Finally, a correlation analysis in 'GPT-3' was calculated between 'in-context learning settings for tasks in GPT-3 (setting)', 'Fluency', 'Feasibility', and 'Originality'; moreover, a correlation analysis in 'Human' was calculated between 'Fluency', 'Feasibility', and 'Originality'.

Table 5 displayed the results of a correlation analysis in 'GPT-3'. The results demonstrated that 'Setting' is weakly related to 'Feasibility' (r=-.324, $p < .01$.), 'Setting' is weakly related to 'Originality' (r=.267, $p < .01$.), 'Fluency' is moderately related to 'Feasibility' (r=-.486, $p < .01$.), 'Fluency' is strongly related to 'Originality' (r=.748, $p < .01$.), and 'Feasibility' is very strongly related to 'Originality' (r=-.807, $p < .01$.). This means that 'Setting' is weakly negative related to 'Feasibility', 'Setting' is weakly related to 'Originality', 'Fluency' is moderately negative related to 'Feasibility', 'Fluency' is strongly related to 'Originality', and 'Feasibility' is very strongly negative related to 'Originality'.

Table 5: **The results of a correlation analysis in 'GPT-3'.** *A number in this table is a correlation coefficient. * indicates a significance level of .05 and ** indicates a significance level of .01. 'Setting' is coded as 0=Zero-shot, 1=One-Shot, and 2=Few-Shot.*

	Setting	Fluency	Feasibility	Originality
Setting	1	.131	-.324**	.267**
Fluency	.131	1	-.486**	.748**
Feasibility	-.324**	-.486**	1	-.807**
Originality	.267**	.748**	-.807**	1

Table 6: **The results of a correlation analysis in 'Human'.** *A number in this table is a correlation coefficient. * indicates a significance level of .05 and ** indicates a significance level of .01.*

	Fluency	Feasibility	Originality
Fluency	1	-.253	.459**
Feasibility	-.253	1	-.508**
Originality	.459**	-.508**	1

Table 6 displayed the results of a correlation analysis in 'Human'. The results displayed that 'Fluency' is moderately related to 'Originality' (r=.459, $p < .01$.), and 'Feasibility' is moderately related to 'Originality' (r=-.508, $p < .01$.). This means that 'Fluency' is moderately related to 'Originality', and 'Feasibility' is moderately negative related to 'Originality'.

5 Discussion and Future Research

This section describes that the results were discussed from the experiments; moreover, the future challenges identified in this study were presented.

5.1 Discussion

First, the text mining results demonstrated that the ideas created by GPT-3 have a high frequency of common words, regardless of the settings. In addition, the ideas created by humans had more

words and parts of speech than those created by GPT-3. The GPT-3 tasks were in Japanese language is considered to contribute to these results. Generally, GPT-3 is pre-trained on a very huge dataset; however, Japanese sentences are less than English sentences in the dataset. Thus, we conclude that the words of the ideas were biased. Therefore, GPT-3 given tasks in Japanese is assumed to be insufficient to support human creative activities.

Next, the comparative results between human and GPT-3 in the ideas evaluation demonstrated that: 1) In fluency and originality, the human-created ideas were evaluated the highest compared to the GPT-3-created ideas, 2) In feasibility, the human-created ideas were evaluated the lowest compared to the GPT-3-created ideas, GPT-3 creates ideas by combining known knowledge; moreover, it does not assume that completely unknown ideas are created. Whereas, humans are possible to create ideas without any constraints. In addition, the task in the experiment contains the phrase 'never been done before', The feasibility is evaluated based on whether the ideas can be implemented using current technology; conversely, the originality is evaluated based on whether the idea has never been done before. Since the fluency is a measure of the appropriateness of an idea for tasks, the correlation analysis of the results showed that; a positive correlation between fluency and originality, as well as a negative correlation between fluency and feasibility. In general, originality weighs in more when assessing creativity [16]. Thus, GPT-3 is lower than human in the qualitative evaluation of ideas. Therefore, GPT-3 is assumed to be utilized as support human creative activities, not be able to create innovative ideas on itself.

And finally, the comparative results among the in-context learning settings ('Setting') for tasks in GPT-3 demonstrated that the ideas with 'Few-Shot' setting was evaluated higher in originality than the ideas with 'Zero-Shot' setting. The correlation analysis of the results demonstrated that; a weakly negative correlation in feasibility, as well as a weakly correlation in originality. Therefore, we conclude that few-shot is the most effective setting for creative activities in the in-context learning for tasks in GPT-3.

5.2 Future research

The author has described the comparison results of the qualitative evaluations in ideas between GPT-3 and humans in this study. Four research questions are proposed that should investigate in the future. First, the GPT-3 tasks were written in Japanese language. Therefore, multi-language tasks in GPT-3 are proposed as experiments of idea creation. Next, the task text given in the GPT-3 leaves room for consideration to customize. The task text in generative AI is available as a common template [17]. Moreover, tuning the instructions given to GPT-3 or asking questions with roles set for GPT-3 are expected to improve the accuracy of responses [18][19]. Then, this study revealed that 'feasibility' and 'originality' are contradictory in the evaluation items. Therefore, 'diversity' is expected to evaluate ideas as a new evaluation item instead of 'feasibility'. And finally, the ideas were evaluated on a four-point scale; however, the problem was revealed that the criteria for the evaluations were different among the judges. Therefore, the evaluations of ideas is proposed the simple choices; 'Yes' or 'No'.

6 Conclusion

This study has described that the qualitative evaluations of ideas have been performed between the GPT3-created ideas and the human-created ideas. In addition, an experiment has been con-

ducted to compare the qualitative evaluations. The evaluation items of ideas have been implemented for three items; fluency, feasibility and originality. In comparison to the GPT3-created ideas and the human-created ideas, the results have demonstrated that; the human-created ideas have been higher than the GPT3-created ideas in fluency and originality, the human-created ideas have been lower than the GPT3-created ideas in feasibility, negative correlations have been detected between fluency and feasibility, and between feasibility and originality, as well as a positive correlation has been between fluency and originality. Furthermore, the results of the in-context learning settings for tasks in GPT-3 have demonstrated that; the few-shot setting has been higher than the zero-shot settings in originality, a weakly negative correlation have been detected in feasibility, as well as a weakly positive correlation in originality. Thus, the conclusions have shown that GPT-3 is lower than human in the qualitative evaluation of ideas, as well as few-shot is the most effective setting for creative activities in the in-context learning for tasks in GPT-3. Therefore, the study has concluded that GPT-3 should be utilized as support human creative activities, not be able to create innovative ideas on itself.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, et al., ”Language models are few-shot learners”, Advances in neural information processing systems, Vol.33, 2020, pp.1877--1901.

[2] R. Dale, “GPT-3: What’s it good for?”, Natural Language Engineering, Vol.27, No.1, Cambridge University Press, 2021, pp.113--118.

[3] L. Floridi and M. Chiriatti, “GPT-3: Its nature, scope, limits, and consequences”, Minds and Machines, Vol.30, No.4, Springer, 2020, pp.681--694.

[4] OpenAI, OpenAI, https://openai.com/ [accessed 2023-10-20].

[5] C. Stevenson, I. Smal, M. Baas R. Grasman and H. van der Maas, “Putting GPT-3’s Creativity to the (Alternative Uses) Test”, arXiv preprint arXiv:2206.08932, 2022.

[6] DeepL, DeepL Translate: The world’s most accurate translator, https://www.deepl.com/translator [accessed 2023-10-20].

[7] H. Furukawa, T. Yuizono, and S. Kunifuji, “Idea Planter: A Backchannel Function for Fostering Ideas in a Distributed Brainstorming Support System”, Proceedings of KICSS’2013, 2013, pp.92--103.

[8] Y. Nagai, T. Taura, and F. Mukai, “Concept blending and dissimilarity: factors for creative concept generation process”, Design studies, Vol.30, No.6, Elsevier, 2009, pp.648--675.

[9] User Local,Inc., User Local AI Text mining Tool, https://textmining.userlocal.jp/ [accessed 2023-10-20].

[10] Y. Hassan-Montero and V. Herrero-Solana, “Improving tag-clouds as visual information retrieval interfaces”, International conference on multidisciplinary information sciences and technologies, Vol.12, 2006.

[11] M.J. Halvey and M.T. Keane, “An assessment of tag presentation techniques”, Proceedings of the 16th international conference on World Wide Web, 2007, pp.1313--1314.

[12] S. Lohmann, J. Ziegler and L. Tetzlaff, “Comparison of tag cloud layouts: Task-related performance and visual exploration”, Human-Computer Interaction--INTERACT 2009: 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part I 12, Springer, 2009, pp.392--404.

[13] R. Vuillemot, T. Clement, C. Plaisant and A. Kumar, “What’s being said near “Martha”? Exploring name entities in literary text collections”, 2009 IEEE Symposium on Visual Analytics Science and Technology, IEEE, 2009, pp.107--114.

[14] F. Heimerl, S. Lohmann, S. Lange and T. Ertl, “Word cloud explorer: Text analytics based on word clouds”, 2014 47th Hawaii international conference on system sciences, IEEE, 2014, pp.1833--1842.

[15] A. Rajaraman and J.D. Ullman, “Mining of massive datasets”, Cambridge University Press, 2011.

[16] J. Diedrich, M. Benedek, E. Jauk, and C.A. Neubauer, “Are creative ideas novel and useful?”, Psychology of aesthetics, creativity, and the arts, Vol.9, No.1, Educational Publishing Foundation, 2015, pp.35--40.

[17] DAIR.AI, Prompt Engineering Guide, https://www.promptingguide.ai/ [accessed 2023-10-20].

[18] J. Wei, M. Bosma, V.Y. Zhao, K. Guu, A.W. Yu, B. Lester, et al., “Finetuned language models are zero-shot learners”, arXiv preprint arXiv:2109.01652, 2021.

[19] T. Kojima, S.S. Gu, M.Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners”, Advances in neural information processing systems, Vol.35, 2022, pp.22199--22213.

[20] M. Takahashi, “Research on Brainstorming (1) : Effectiveness of “Rules of Idea gene-ra-tion””, Journal of Japan Creativity Society, No.2, 1998, pp.94--122 (in Japanese).