Building Japanese Creativity Benchmarks and Applying them to Enhance LLM Creativity

Anonymous ACL submission

Abstract

To evaluate the creativity of large language models (LLMs) in Japanese, we construct three benchmarks: Japanese Creativity Questions (JCQ), Divergent Association Task (DAT), and Story Alteration Task (SAT). JCQ comprehensively evaluates creativity using LLMs. Meanwhile, DAT and SAT measure specific aspects of creative ability using embeddings. We also analyze correlations between JCQ and DAT, JCQ and SAT, and DAT and SAT. While JCQ provides comprehensive evaluation, it is relatively time and resource intensive. In contrast, DAT and SAT offer lower comprehensiveness but enable quick, low-cost assessment. Additionally, we investigate whether training with DAT contributes to enhancing LLM creativity.

1 Introduction

Creativity is a crucial ability that has supported human progress and development. Creative thinking has been central to human activities, from artistic expression and scientific discovery to solving social problems. In recent years, with the development of large language models (LLMs), AI systems have shown potential to support and extend human creative activities in text generation and problem-solving, leading to active research in this area (Franceschelli and Musolesi, 2024; Tanaka et al., 2024; Watanabe et al., 2024; Li et al., 2024). For both humans and LLMs, creativity has become an essential element for addressing the challenges of our increasingly complex society and creating new value.

Previous research on LLM creativity has primarily focused on English, but there are differences in how creativity manifests and is evaluated across languages and cultures. Japanese, in particular, has different grammatical structures and expressive styles from English, with unique linguistic characteristics such as abundant homonyms and high context-dependency. These characteristics may uniquely influence LLMs' creative expression, highlighting the importance of cross-linguistic creativity research.

In this study, we construct three benchmarks to measure LLM creativity in Japanese either comprehensively or efficiently depending on the purpose, and evaluate several LLMs. The first is Japanese Creativity Questions (JCQ), developed based on the verbal tasks of the Torrance Test of Creative Thinking (TTCT) (Torrance, 1966), which is widely used to evaluate human creativity. This follows the approach of previous research (Zhao et al., 2024). It consists of seven tasks and uses four criteria for evaluation. The second is the Divergent Association Task (DAT) (Olson et al., 2021), which requires listing words that are as semantically distant from each other as possible. The third is the Story Alteration Task (SAT), which measures how much a story differs from the original after being altered. JCQ evaluation uses a powerful LLM as LLM-as-a-judge, while DAT and SAT evaluations use embeddings. JCQ can comprehensively evaluate creativity but requires time and resources for assessment. DAT and SAT, on the other hand, can quickly and easily measure specific aspects of creativity by using embeddings. This allows for choosing between comprehensive or rapid evaluation methods to measure LLM creativity according to specific needs.

Furthermore, we investigate whether training LLMs using DAT improves creativity through generalization ability, potentially enhancing scores on JCQ and SAT.

2 Related Work

The Torrance Test of Creative Thinking (TTCT) is widely known as a test for evaluating human creativity. It consists of verbal and figural tests with free-response questions, such as "List as many unusual uses for a light bulb as possible." When evaluating responses, four criteria are commonly

Task	Definition	Example Question (Translated)
Unusual Uses	A task to think of unusual or diverse uses for common objects.	Please list as many unusual uses for a light bulb as possible.
Consequences	A task to predict consequences or impacts in unusual or hypothetical situations.	What would be the effects on society and daily life if the internet became unavailable worldwide for 24 hours?
Just Suppose	A task to consider hypothetical, often fantastical scenarios and their implications.	You have gained the power to make objects disappear. What would you eliminate? Please list as many ideas as possible.
Situation	A task to respond to a given situation.	If gravity were to reverse direction, how would you survive on the ground?
Common Problem	A task to generate solutions to problems that are familiar and everyday for most people.	Please suggest ways to efficiently manage the contents of a refrigerator.
Improvement	A task to improve or modify existing objects or ideas.	Please list as many ways as possible to make a standard bed more comfortable.
Imaginative Stories	A task to create a story with a given prompt.	Please create a story with the title "The Library on the Far Side of the Moon"

Table 1: Definitions and example questions for JCQ tasks. Created with reference to previous research (Zhao et al., 2024).

Criterion	Definition
Fluency	The ability to generate numerous relevant ideas in response to a given question. Essentially measures the quantity of ideas.
Flexibility	The diversity of categories from which ideas can be generated. The ability to think of alternatives, shift from one class or perspective to another, or approach a given problem or task from various angles.
Originality	The uniqueness of the ideas generated. Unique ideas are those that are unusual, rare, or unconventional.
Elaboration	The ability to develop, refine, and embellish ideas. Includes adding details, developing nuances, and making basic concepts more intricate or complex.

Table 2: Definitions of the four criteria in JCQ. Following previous research (Zhao et al., 2024).

used: Fluency, Flexibility, Originality, and Elaboration. These four criteria are generally adopted in many other creativity studies (Lu et al., 2024; Handayani et al., 2021; Hong et al., 2013). TTCT is widely used in the field of psychology and is considered an excellent test that can measure the creativity of many people (Kim, 2006).

The Divergent Association Task (DAT) has also been developed as a creativity test, with research conducted on human subjects (Olson et al., 2021). DAT is a task to list words that are as semantically distant from each other as possible, with higher scores awarded for greater semantic distances between words. They also conducted the Alternative Uses Task (AUT), which asks participants to list as many uses as possible for common objects like "newspaper" or "shoe." Their results showed significant correlations between DAT scores and Flexibility and Originality scores in AUT.

In English, there is a study that created tests based on the verbal tests of TTCT and measured LLM creativity using OpenAI's GPT-4 as an evaluator (Zhao et al., 2024). However, in Japanese, benchmarks for evaluating LLM creativity are not currently known.

Regarding the enhancement of human creativity, training with verbal divergent thinking exercises has been shown to improve specific aspects of creativity (Fink et al., 2015). For enhancing LLM creativity, prompting strategies that promote associative thinking—the cognitive process of connecting unrelated concepts—have been found to improve certain aspects of creativity (Mehrotra et al., 2024).

3 Construction of Japanese Creativity Benchmarks

We construct three benchmarks to facilitate either comprehensive or efficient assessment of LLM creativity in Japanese, depending on the evaluation purpose.

3.1 Japanese Creativity Questions (JCQ)

JCQ was created following previous research (Zhao et al., 2024) with the aim of comprehensively measuring creativity. Through conversations with OpenAI's GPT-4o, o1-preview, and Anthropic's Claude 3.5 Sonnet, we created 100 questions for each of the seven tasks used in Zhao et al. (2024), for a total of 700 Japanese questions. The task definitions and example questions are shown in Table 1. An example LLM response is shown in Table 15 in the

	Fluency	Flexibility	Originality	Elaboration	Mean
GPT-4o	4.10	**4.28**	2.73	3.47	3.64
Claude 3.5 Sonnet	**4.29**	4.04	2.73	2.87	3.48
calm3-22b	4.16	4.18	**2.87**	**3.86**	**3.76**
llm-jp-3-13b	3.74	3.79	2.65	3.45	3.41
Swallow-8B	3.91	3.45	2.34	2.79	3.12

Table 3: Mean scores across all tasks for each model and criterion in JCQ.

	Unusual Uses	Consequences	Just Suppose	Situation	Common Problem	Improvement	Imaginative Stories
GPT-4o	**3.97**	3.69	3.83	3.28	3.48	**4.01**	3.25
Claude 3.5 Sonnet	3.73	3.42	3.80	3.08	**3.61**	3.80	2.93
calm3-22b	3.84	3.92	**3.91**	**3.73**	3.45	4.00	**3.50**
llm-jp-3-13b	3.08	**3.92**	3.52	3.69	3.00	3.64	3.01
Swallow-8B	3.28	3.33	3.39	2.80	3.08	3.45	2.54

Table 4: Mean scores across all criteria for each model and task in JCQ.

appendix.

Evaluation is conducted using LLM-as-a-Judge, the effectiveness of which has already been demonstrated (Zheng et al., 2023). Specifically, model responses are evaluated on a scale of 1 to 5 across four criteria: Fluency, Flexibility, Originality, and Elaboration. Each criterion is defined as shown in Table 2, following Zhao et al. (2024).

3.2 Divergent Association Task (DAT)

DAT is a test used in previous research (Olson et al., 2021) that requires listing 10 words that are as semantically distant from each other as possible. Higher creativity is indicated by more semantically distant words. This test was developed to measure human creativity, but our study targets LLMs. An example LLM response is shown in Table 16 in the appendix.

The evaluation uses embeddings of each of the 10 words listed by the model. The score for one trial is the mean of the cosine distances ($1 -$ cosine similarity) between all pairs of words. Multiple trials are conducted, and the mean score across these trials becomes the model's score.

3.3 Story Alteration Task (SAT)

SAT, proposed in this paper, is a test that involves rewriting stories according to specific instructions. Higher creativity is indicated by greater differences between the rewritten story and the original. An example response is shown in Table 17 in the appendix.

The evaluation uses embeddings of the original story and the story output by the model. The cosine distance between the two embeddings is calculated, and the mean across multiple stories becomes the model's score.

4 Creativity Evaluation Experiments for LLMs

We evaluate the creativity of five LLMs using the three constructed benchmarks.

4.1 Experimental Setup

We have the following five models generate responses. The temperature is set to 1.

- gpt-4o-2024-08-06[1] (GPT-4o)
- claude-3-5-sonnet-20241022[2] (Claude 3.5 Sonnet)
- calm3-22b-chat[3] (calm3-22b)
- llm-jp-3-13b-instruct[4] (llm-jp-3-13b)
- Llama-3.1-Swallow-8B-Instruct-v0.1[5] (Swallow-8B)

For JCQ, we use GPT-4o for evaluation. The evaluation prompt is shown in Table 20 in the appendix.

For DAT, we set the number of trials to calculate the model's mean score to 100. Responses that do not follow the specified format or contain non-Japanese words, symbols, or non-nouns are excluded from evaluation and not counted in

[1] https://platform.openai.com/docs/models#gpt-4o

[2] https://docs.anthropic.com/en/docs/about-claude/models#model-names

[3] https://huggingface.co/cyberagent/calm3-22b-chat

[4] https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

[5] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1

	Fluency	Flexibility	Originality	Elaboration	Mean
Unusual Uses	4.50	4.13	2.92	2.78	3.58
Consequences	4.00	4.31	2.67	3.64	3.65
Just Suppose	4.58	4.43	2.64	3.11	3.69
Situation	3.30	4.03	2.57	3.38	3.32
Common Problem	3.98	3.85	2.01	3.46	3.32
Improvement	4.71	4.51	2.72	3.17	3.78
Imaginative Stories	3.22	2.36	3.12	3.49	3.05

Table 5: Mean scores across all models for each task and criterion in JCQ.

	Score	Std.
GPT-4o	0.527	0.014
Claude 3.5 Sonnet	0.530	0.018
calm3-22b	0.514	0.018
llm-jp-3-13b	0.494	0.049
Swallow-8B	0.505	0.014

Table 6: Results of DAT.

	Score
GPT-4o	0.526
Claude 3.5 Sonnet	0.579
calm3-22b	0.458
llm-jp-3-13b	0.219
Swallow-8B	0.193

Table 7: Results of SAT.

the number of trials. We use the Japanese morphological analyzer Juman++[6] for noun validation. The prompt is shown in Table 18 in the appendix. For the embedding model for evaluation, we use GLuCoSE-base-ja-v2[7].

For SAT, we use 113 fairy tales as original stories. These are fairy tales selected from a fairy tale website[8] and summarized to approximately 200-400 characters using gpt-4o-2024-05-13[1]. The rewriting instruction is to transform the fairy tale into a modern-style story. The prompt is shown in Table 19 in the appendix. For the embedding model for evaluation, we use simcse-ja-bert-base-clcmlp[9]. We choose this model because it has a high correlation with human creativity evaluations. For details, please refer to Section C.2 in the appendix.

4.2 Results

4.2.1 Japanese Creativity Questions (JCQ)

The mean scores across all tasks for each model and criterion are shown in Table 3. There were characteristics such as larger differences in Elaboration scores between models compared to differences in Fluency and Originality.

The mean scores across all criteria for each model and task are shown in Table 4. Overall, there were characteristics such as models performing well on the Improvement task and struggling with the Imaginative Stories task.

The mean scores across all models for each task and criterion are shown in Table 5. There were characteristics such as notably low Flexibility in the Imaginative Stories task and low Originality in the Common Problem task compared to other tasks.

4.2.2 Divergent Association Task (DAT)

The scores for each model are shown in Table 6. The two models considered powerful, GPT-4o and Claude 3.5 Sonnet, achieved high scores.

4.2.3 Story Alteration Task (SAT)

The scores for each model are shown in Table 7. Claude 3.5 Sonnet's score was notably high. The second highest score was achieved by GPT-4o, indicating that, similar to DAT, the two models considered powerful performed well.

4.3 Analysis

4.3.1 Correlation between GPT-4o and Human Evaluation in JCQ

Some responses to JCQ were manually evaluated. Three university students collaboratively evaluated 15 responses for each task, totaling 105 responses, using the same method as GPT-4o. The Pearson correlation with GPT-4o's evaluation is shown in Table 8. We calculated the correlation between GPT-4o and human evaluation scores for each task and criterion in JCQ. Overall, there was correlation, but some tasks and criteria showed weak correlation. In particular, the correlation was weak for the Imaginative Stories task. This suggests that GPT-4o, may not effectively evaluate the creativity of stories like humans.

[6] https://github.com/ku-nlp/jumanpp
[7] https://huggingface.co/pkshatech/GLuCoSE-base-ja-v2
[8] https://www.douwa-douyou.jp/index.shtml
[9] https://huggingface.co/pkshatech/simcse-ja-bert-base-clcmlp

	Fluency	Flexibility	Originality	Elaboration	Mean
Unusual Uses	**1.000**	0.222	0.208	**0.613**	**0.570**
Consequences	**0.688**	**0.668**	**0.696**	**0.745**	**0.791**
Just Suppose	**0.964**	**0.623**	**0.733**	**0.683**	**0.755**
Situation	0.299	**0.619**	**0.551**	0.174	**0.707**
Common Problem	**0.814**	**0.640**	**0.539**	0.494	**0.639**
Improvement	**0.868**	**0.552**	0.346	**0.730**	0.426
Imaginative Stories	0.488	0.340	-0.213	-0.076	0.397
All	**0.683**	**0.577**	**0.525**	**0.546**	**0.654**

Table 8: Correlation between GPT-4o and human evaluation scores for each task and criterion in JCQ. Bold values indicate p-values below 0.05.

	Fluency	Flexibility	Originality	Elaboration	Mean
Unusual Uses	0.847	**0.952**	0.455	-0.037	**0.883**
Consequences	-0.154	-0.308	-0.118	-0.316	-0.340
Just Suppose	**0.890**	0.819	0.567	-0.058	0.722
Situation	-0.549	0.063	-0.035	-0.447	-0.290
Common Problem	0.825	**0.933**	0.329	0.335	**0.948**
Improvement	0.844	0.848	0.755	-0.469	0.633
Imaginative Stories	0.046	-0.042	0.826	0.512	0.287
All	**0.916**	0.670	0.437	-0.108	0.466

Table 9: Correlation between JCQ and DAT. The table shows the correlation between model scores for each task and criterion in JCQ and the model scores in DAT. Bold values indicate p-values below 0.05.

4.3.2 Correlation between JCQ and DAT

The Pearson correlation between JCQ and DAT is shown in Table 9. We calculated the correlation between model scores for each task and criterion in JCQ and the model scores in DAT. Strong correlations were found in Fluency and Flexibility for some tasks. In particular, there was a strong correlation between Flexibility in the Unusual Uses task and DAT, which aligns with previous research on humans (Olson et al., 2021) that found a correlation between Flexibility in AUT (a task similar to Unusual Uses) and DAT. However, while that research found a correlation between Originality in AUT and DAT for humans, our study found a weak correlation between Originality in the Unusual Uses task and DAT for LLMs. This suggests that correlation patterns between tasks may not always be consistent between LLMs and humans.

4.3.3 Correlation between JCQ and SAT

The Pearson correlation between JCQ and SAT is shown in Table 10. We calculated the correlation between model scores for each task and criterion in JCQ and the model scores in SAT. Strong correlations were found in Flexibility and Originality for some tasks, and overall, the correlation with JCQ was stronger than with DAT.

4.3.4 Correlation between DAT and SAT

The Pearson correlation between DAT and SAT was 0.933, with a p-value of 0.021. The strong correlation likely stems from the fact that both tasks award higher scores when the generated text is semantically distant from the context.

5 Training LLMs using DAT

We investigate whether using DAT, which promotes divergent thinking, as training data can effectively enhance LLM creativity. Since DAT measures the ability to generate semantically distant words, it is suitable for training the ability to form new connections between concepts—an important aspect of creativity. We examine whether this training affects not only DAT scores themselves but also scores on more comprehensive creativity measures such as JCQ and SAT.

5.1 Method

We separately perform three distinct training approaches: SFT (Ouyang et al., 2022), DPO (Rafailov et al., 2024), and GRPO (Shao et al., 2024) using DAT on the following three models:

- Llama-3.1-Swallow-8B-Instruct-v0.3[10] (Swallow-8B)
- Qwen2.5-7B-Instruct[11] (Qwen-2.5-7B)

[10] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3

[11] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

	Fluency	Flexibility	Originality	Elaboration	Mean
Unusual Uses	0.606	**0.992**	0.736	0.114	**0.899**
Consequences	0.126	-0.200	0.214	-0.076	-0.017
Just Suppose	0.678	**0.945**	0.824	0.260	**0.897**
Situation	-0.221	0.368	0.320	-0.117	0.058
Common Problem	0.627	**0.978**	0.625	0.573	**0.981**
Improvement	0.601	**0.966**	**0.939**	-0.230	0.812
Imaginative Stories	0.331	0.237	**0.960**	0.741	0.556
All	**0.908**	0.855	0.725	0.170	0.712

Table 10: Correlation between JCQ and SAT. The table shows the correlation between model scores for each task and criterion in JCQ and the model scores in SAT. Bold values indicate p-values below 0.05.

		Valid Responses	Mean	Std.	Unique Words
Random		131072	0.555	0.020	22085
Swallow-8B		105401	0.524	0.018	8026
	SFT	100991	0.538	0.022	17614
	DPO 1	129447	0.547	0.017	7231
	DPO 2	130450	0.594	0.014	5689
	GRPO	117824	0.570	0.022	10696
Qwen2.5-7B		81548	0.519	0.020	7839
	SFT	81772	0.526	0.023	13470
	DPO 1	112768	0.536	0.015	5949
	DPO 2	115034	0.554	0.015	4464
	GRPO	96567	0.541	0.022	8431
llm-jp-3-7.2b		25556	0.521	0.040	20999
	SFT	48830	0.534	0.039	41410
	DPO 1	123998	0.533	0.024	25782
	DPO 2	127845	0.567	0.019	16420
	GRPO	14668	0.558	0.026	30548

Table 11: Results of DAT training. The table shows the number of valid responses, mean score, standard deviation, and number of unique words before and after training. The values are aggregated for valid responses (those with non-zero scores) out of 131,072 responses.

- llm-jp-3-7.2b-instruct2[12] (llm-jp-3-7.2b)

5.1.1 SFT

We implement DAT-based SFT within the instruction tuning framework. The training data consists of the top 16,384 scoring responses from 131,072 DAT responses created using random words. DAT scores are calculated using the mean cosine distance between embeddings of generated words, as described in Section 4. Random words are obtained from a noun list created from the dictionary of the Japanese morphological analyzer Juman++[13]. We train for one epoch with a learning rate of 2e-7 and a batch size of 256.

5.1.2 DPO

The training data consists of the top 16,384 scoring responses from 131,072 responses generated by the model itself as "chosen" and the bottom 16,384 as "rejected." Responses that do not follow the format or contain non-Japanese words, symbols, or non-nouns are not excluded but given a score of 0. We train for one epoch with a learning rate of 5e-7 and a batch size of 256. Additionally, we create new training data using the trained model and perform a second stage of training.

5.1.3 GRPO

The reward is set to 10 times the DAT score. Responses that do not follow the format or contain non-Japanese words, symbols, or non-nouns receive a reward of 0. Responses identical to previous ones also receive a reward of 0. We train for one epoch with 4,096 training samples, 8 generations, a learning rate of 5e-7, and a batch size of 256.

5.2 Results

The results of DAT training are shown in Table 11. The table shows the number of valid responses, mean DAT score, standard deviation, and number of unique words before and after training. The values are aggregated for valid responses (those with non-zero scores) out of 131,072 responses.

[12] https://huggingface.co/llm-jp/llm-jp-3-7.2b-instruct2

[13] https://github.com/ku-nlp/JumanDIC/blob/master/dic/ContentW.dic

		Fluency	Flexibility	Originality	Elaboration	Mean
Swallow-8B		4.52	3.78	2.85	3.61	3.69
	SFT	4.51	3.76	2.86	3.64	3.69
	DPO 1	4.54	3.76	2.83	3.62	3.69
	DPO 2	4.51	3.70	2.86	3.60	3.67
	GRPO	4.52	3.73	2.87	3.60	3.68
Qwen2.5-7B		4.05	3.92	2.88	2.98	3.46
	SFT	4.05	3.91	2.87	2.93	3.44
	DPO 1	4.09	3.94	2.91	3.00	3.48
	DPO 2	4.06	3.95	2.85	3.02	3.47
	GRPO	4.02	3.94	2.90	3.00	3.47
llm-jp-3-7.2b		3.77	3.81	2.66	3.42	3.42
	SFT	3.79	3.81	2.65	3.38	3.41
	DPO 1	3.83	3.78	2.66	3.40	3.42
	DPO 2	3.92	3.87	2.69	3.46	3.48
	GRPO	3.64	3.68	2.64	3.29	3.31

Table 12: Mean scores across all tasks for each model and criterion in JCQ for models trained with DAT.

		Unusual Uses	Consequences	Just Suppose	Situation	Common Problem	Improvement	Imaginative Stories
Swallow-8B		3.72	3.94	3.76	3.38	3.87	3.96	3.20
	SFT	3.71	3.91	3.79	3.36	3.90	3.92	3.24
	DPO 1	3.67	3.92	3.78	3.40	3.88	3.92	3.25
	DPO 2	3.64	3.93	3.74	3.39	3.86	3.90	3.23
	GRPO	3.68	3.93	3.76	3.41	3.82	3.92	3.23
Qwen2.5-7B		3.54	3.84	3.53	3.28	3.18	3.81	3.04
	SFT	3.52	3.82	3.50	3.28	3.15	3.82	3.00
	DPO 1	3.62	3.84	3.50	3.30	3.21	3.74	3.18
	DPO 2	3.57	3.79	3.55	3.25	3.19	3.81	3.11
	GRPO	3.59	3.78	3.60	3.34	3.14	3.80	3.00
llm-jp-3-7.2b		3.09	3.84	3.68	3.76	3.01	3.19	3.36
	SFT	3.02	3.83	3.74	3.72	2.97	3.23	3.34
	DPO 1	3.19	3.82	3.64	3.73	2.93	3.31	3.32
	DPO 2	3.38	3.87	3.68	3.78	2.93	3.36	3.40
	GRPO	2.74	3.80	3.50	3.76	2.92	3.12	3.36

Table 13: Mean scores across all criteria for each model and task in JCQ for models trained with DAT.

		Score
Swallow-8B		0.421
	SFT	0.431
	DPO 1	0.430
	DPO 2	0.410
	GRPO	0.417
Qwen2.5-7B		0.450
	SFT	0.435
	DPO 1	0.447
	DPO 2	0.439
	GRPO	0.454
llm-jp-3-7.2b		0.185
	SFT	0.179
	DPO 1	0.172
	DPO 2	0.140
	GRPO	0.210

Table 14: Mean scores in SAT for models trained with DAT.

The two-stage DPO showed the largest increase in score. The ratio of unique words to valid responses increased with SFT and decreased with DPO.

The mean scores across all tasks for each model and criterion in JCQ for models trained with DAT are shown in Table 12. In most cases across training methods and criteria, scores hardly increased from the original model. As an exception, the Fluency score improved when llm-jp-3-7.2b was trained with DPO.

The mean scores across all criteria for each model and task in JCQ for models trained with DAT are shown in Table 13. In most cases across training methods and tasks, scores hardly increased from the original model. As an exception, the Unusual Uses and Improvement task scores improved when llm-jp-3-7.2b was trained with DPO.

The mean scores in SAT for models trained with DAT are shown in Table 14. In most cases across training methods, scores hardly increased from the original model. As an exception, the score improved when llm-jp-3-7.2b was trained with GRPO.

5.3 Discussion

The model with the most unique words in DAT was llm-jp-3-7.2b. This is likely because this model was trained on a large Japanese corpus and uses a tokenizer extended for Japanese.

The increase in the ratio of unique words to valid responses with SFT is likely because the training data contained many new words that the original model did not generate. Conversely, the decrease with DPO is likely because the training led to an increased probability of generating responses using specific groups of words that yield high scores.

There are several possible reasons why the Fluency, Unusual Uses, and Improvement scores for llm-jp-3-7.2b improved in JCQ after DAT training. First, this model initially had few valid responses in DAT. The increase in valid responses through training may have improved instruction following, thereby improving JCQ scores. Additionally, DAT training may have enhanced the ability to enumerate items, improving scores on the criterion that measures the quantity of ideas and the tasks that require enumeration. The model's extensive training in Japanese and use of a tokenizer extended for Japanese may also be factors.

6 Conclusion

We constructed three benchmarks to measure LLM creativity: JCQ, DAT, and SAT. Each benchmark has advantages and disadvantages in terms of comprehensiveness and ease of use. JCQ uses seven tasks and four criteria, allowing for comprehensive creativity evaluation, but requires more time and resources compared to the other two benchmarks as it uses LLMs for evaluation. DAT has low comprehensiveness with only one prompt but allows for rapid evaluation using embeddings. SAT requires preparing original stories but enables easy evaluation using embeddings. Its comprehensiveness is lower than JCQ as it involves only one task of rewriting stories, but higher than DAT as it uses multiple stories.

We also analyzed the correlation between GPT-4o and human evaluation in JCQ. Overall, there was correlation except for some tasks and criteria, particularly the Imaginative Stories task. This suggests that JCQ results are reliable except for the weakly correlated parts.

Furthermore, we analyzed correlations between JCQ and DAT, JCQ and SAT, and DAT and SAT. DAT and SAT correlated with JCQ in some tasks and criteria, with SAT showing stronger correlation with JCQ overall. This indicates a trade-off between ease of use and strength of correlation with JCQ, as DAT is easier to use than SAT. DAT and SAT showed strong correlation with each other, possibly due to similarities in task nature.

We also investigated whether DAT training improves creativity through generalization ability, potentially enhancing JCQ and SAT scores. While scores generally did not increase, there were cases where scores improved under specific conditions.

Properly evaluating creativity is important for understanding and utilizing LLM capabilities. This study proposes an initial framework for evaluating LLM creativity in Japanese. The three proposed benchmarks provide means to efficiently measure LLM creativity according to purpose. This enables understanding the current state of LLMs' creative abilities and selecting appropriate models for specific tasks and applications.

Future challenges include establishing more refined approaches for creativity evaluation. In particular, developing evaluation methods that consider Japanese-specific linguistic and cultural characteristics, and improving methodologies to enhance consistency with human evaluation are needed. Exploring effective training methods to enhance creativity is also an important research direction. Through such efforts, we can expect improvements in LLMs' creative abilities and the development of appropriate evaluation methods.

Limitations

Our study has several limitations. First, while JCQ provides comprehensive creativity evaluation, GPT-4o's evaluations showed weak correlation with human judgments for certain tasks, particularly Imaginative Stories. This suggests that LLM-as-a-judge approaches may not fully capture human perceptions of creativity in narrative contexts.

Second, DAT and SAT, though efficient, measure only specific aspects of creativity—semantic distance between words and story rewriting ability, respectively. They cannot capture the full spectrum of creative capabilities that JCQ attempts to measure.

Finally, our experiments with DAT-based training showed few improvements in other creativity tests. While specific scores improved under certain conditions (e.g., llm-jp-3-7.2b's Fluency after DPO training), the overall lack of consistent improvements suggests that training specifically on semantic distance tasks may not generalize well to broader creative abilities. More sophisticated training approaches that target multiple aspects of creativity simultaneously may be necessary for meaningful enhancement of LLM creative capabilities.

References

Andreas Fink, Mathias Benedek, Karl Koschutnig, Eva Pirker, Elisabeth Berger, Sabrina Meister, Aljoscha C. Neubauer, Ilona Papousek, and Elisabeth M. Weiss. 2015. Training of verbal creativity modulates brain activity in regions associated with language- and memory-related demands. *Human brain mapping*, 36(10):4104–4115.

Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI & society*.

S A Handayani, Y S Rahayu, and R Agustini. 2021. Students' creative thinking skills in biology learning: fluency, flexibility, originality, and elaboration. *Journal of Physics: Conference Series*, 1747(1):12040–.

Eunsook Hong, Yun Peng, Harold F. O'Neil Jr, and Junbin Wu. 2013. Domain-general and domain-specific creative-thinking tests: Effects of gender and item content on test performance. *The Journal of creative behavior*, 47(2):89–105.

Kyung Hee Kim. 2006. Can we trust creativity tests? a review of the torrance tests of creative thinking (ttct). *Creativity Research Journal*, 18(1).

Jiayang Li, Jiale Li, Yunsheng Su, Stavroula Ntoa, Helmut Degen, Stavroula Ntoa, and Helmut Degen. 2024. A map of exploring human interaction patterns with llm: Insights into collaboration and creativity. In *Artificial Intelligence in HCI*, volume 14736 of *Lecture Notes in Computer Science*, pages 60–85. Springer, Switzerland.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. arXiv. Abs/2401.12491.

Pronita Mehrotra, Aishni Parab, and Sumit Gulwani. 2024. Enhancing creativity in large language models through associative thinking strategies. arXiv. Abs/2405.06715.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv. Abs/2203.02155.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. arXiv. Abs/2305.18290.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv. Abs/2402.03300.

Takaaki Tanaka, Shun Otsubo, Kotaro Ito, Takuya Hatakeyama, Yuji Anzai, Tomoaki Nagasaka, Takashi Matsui, and Nobuyuki Ishikawa. 2024. Research on ideation applications using llm-based multi-agent systems and idea evaluation methods. *Proceedings of the Annual Conference of JSAI*, pages 4G3GS205–4G3GS205.

E. P. Torrance. 1966. *Torrance tests of creative thinking: Directions manual and scoring guide*. Personnel Press.

Kengo Watanabe, Takashi Kawamura, Reo Kobayashi, Kzuma Ari, Akifumi Ito, and Satoshi Kurihara. 2024. Interactive story generation system: Enhancing creative writing with a llm informed by narrative structure analysis. *Proceedings of the Annual Conference of JSAI*, pages 1T3OS32a05–1T3OS32a05.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, Qi Guo, Ling Li, and Yunji Chen. 2024. Assessing and understanding creativity in large language models. arXiv. Abs/2401.12491.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. arXiv. Abs/2401.12491.

A Example LLM Responses

Question (Translated)	Please list as many unusual uses for a light bulb as possible.
Model	Claude 3.5 Sonnet
Response (Translated)	Let me suggest some creative uses for a light bulb: Decoration & Crafts 1. Miniature terrarium container 2. Christmas ornament 3. Single flower vase 4. Pen holder 5. Candle holder 6. Art object 7. Doorknob 8. Small item container Gardening Related 9. Container for growing houseplants 10. Seed starting pot 11. Hydroponic container 12. Mini greenhouse Practical Uses 13. Salt and pepper shaker 14. Spice container 15. Stationery holder 16. Piggy bank 17. Moisture-proof case 18. Temporary medicine storage 19. Pin and clip holder 20. Small aquarium Crafts & Science Experiments 21. Material for making snowmen 22. Science experiment equipment 23. Doorbell 24. Material for an hourglass 25. Material for wind chimes (continued)

Table 15: Example response for JCQ.

Model	Claude 3.5 Sonnet
Response (Translated)	1. Umbrella 2. Sugar 3. Map 4. Music 5. Battery 6. Mirror 7. Air 8. Clock 9. Fireworks 10. Newspaper

Table 16: Example response for DAT.

Prompt (Translated)	Please rewrite the following story according to the conditions. # Conditions - Create a creative modern-style story - About 400 characters in length - No line breaks - Output only the story you created # Story Long ago, a kind-hearted fisherman named Urashima Taro saved a turtle that was being bullied by children. A few days later, the turtle invited Taro to the Dragon Palace as a token of gratitude, where he spent dreamlike days with a beautiful princess, luxurious food, and dancing fish. After a few days, when Taro decided to return home, the princess gave him a box that he must never open. When Taro returned to his village, everything had changed, and he couldn't find his house or mother. After wandering in confusion, he opened the box, and with white smoke, he instantly became an old man. The joyful days at the Dragon Palace had actually spanned hundreds of years.
Model	Claude 3.5 Sonnet
Response (Translated)	Takuya Urashima, a software engineer at a major IT company, witnessed a malfunctioning self-driving car on the roadside. Inside was a distressed elderly woman, and Takuya used his programming skills to repair the car's system. Days later, he received an invitation to "Ryugu," a luxury resort hotel, from the elderly woman. There, he met her granddaughter Otohime, a VR technology genius, and Takuya was captivated by the latest metaverse world she had developed. While immersed in the highly realistic virtual world, Takuya decided to return to reality. Upon returning home wearing the latest smartwatch Otohime had given him at their parting, he discovered that three years had passed. When the panicked Takuya operated the watch, the facial recognition system suddenly activated. The face reflected there was his own, rapidly aged by stress. The happy time in the metaverse had actually been a long period of social withdrawal in reality.

Table 17: Example response for SAT.

B Instruction and Evaluation Prompts

```
# Instructions
Think of 10 words that have meanings or uses as different from each other as possible, and output them in the following format.

# Rules
1. Each word should be a single word only
2. Only nouns are allowed (objects, entities, concepts, etc.)
3. Proper nouns cannot be used (specific people, places, etc.)
4. Technical terms cannot be used
5. No explanations are needed

# Output Format
1. word1
2. word2
3. word3
4. word4
5. word5
6. word6
7. word7
8. word8
9. word9
10. word10
```

Table 18: Prompt for DAT (Translated). Created with reference to Olson et al. (2021).

```
Please rewrite the following story according to the conditions.
# Conditions
- Create a creative modern-style story
- About 400 characters in length
- No line breaks
- Output only the story you created
# Story
{original story}
```

Table 19: Prompt for SAT (Translated).

Read the response to the question and evaluate it on a 5-point scale from four perspectives.

Notes
- Read the entire response
- Read the explanation for each criterion carefully and evaluate independently
- If you are unsure about the evaluation, choose the lower rating
- Follow the output format and output only the evaluation results

Output Format
Fluency: [1-5]
Flexibility: [1-5]
Originality: [1-5]
Elaboration: [1-5]

Question
{question}

Response
{response}

Fluency: Evaluate the number of different ideas related to the question. Count repetitions or paraphrases as a single idea.
1. 1-2 ideas
2. 3-4 ideas
3. 5-6 ideas
4. 7-8 ideas
5. 9 or more ideas

Flexibility: Evaluate the diversity of perspectives, categories, or approaches shown in the response.
1. Single perspective
2. 2 different perspectives
3. 3 different perspectives
4. 4 different perspectives
5. 5 or more different perspectives

Originality: Evaluate how unique the ideas in the response are.
1. Extremely common ideas that anyone would think of
2. Common ideas with slight innovation
3. Somewhat unusual ideas with elements of surprise
4. Novel and original ideas
5. Extremely unique and innovative ideas

Elaboration: Evaluate the detail and depth of idea development.
1. Ideas are simple with no detailed explanation
2. Basic explanations are included but no deep development
3. Some detailed explanations or developments
4. Ideas are explained in detail and well developed
5. Ideas are very detailed with complex developments

Table 20: Evaluation prompt for JCQ (Translated).

C Detailed SAT Experiment

We conduct SAT experiments on the following 11 models:

- gpt-4o-2024-05-13[1] (GPT-4o)
- gpt-4-turbo-2024-04-09[14] (GPT-4 Turbo)
- gpt-3.5-turbo-0125[15] (GPT-3.5 Turbo)
- claude-3-5-sonnet-20240620[2] (Claude 3.5 Sonnet)
- claude-3-opus-20240229[2] (Claude 3 Opus)
- claude-3-sonnet-20240229[2] (Claude 3 Sonnet)
- claude-3-haiku-20240307[2] (Claude 3 Haiku)
- Meta-Llama-3-70B-Instruct[16] (Llama-3-70B)
- Meta-Llama-3-8B-Instruct[17] (Llama-3-8B)
- Qwen2-72B-Instruct[18] (Qwen2-72B)
- Qwen2-7B-Instruct[19] (Qwen2-7B)

In addition to evaluation using the simcse-ja-bert-base-clcmlp embedding model, we also conduct human evaluation and GPT-4o evaluation.

Human evaluation is performed via crowdsourcing. Crowdworkers are presented with the original story and 11 stories generated by the models, and asked to rank them in order of perceived creativity. Scores are assigned from 1 point for first place, 0.9 points for second place, 0.8 points for third place, and so on down to 0 points, with the model's score being the mean across all stories. The evaluation instructions for crowdworkers are shown in Table 21.

For GPT-4o evaluation, we present the original story and the story generated by the model, and evaluate creativity on a scale of 1 to 5. The model's score is the mean across all stories divided by 5. The evaluation prompt is shown in Table 22.

[14] https://platform.openai.com/docs/models/#gpt-4-turbo-and-gpt-4

[15] https://platform.openai.com/docs/models/#gpt-3-5-turbo

[16] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

[17] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[18] https://huggingface.co/Qwen/Qwen2-72B-Instruct

[19] https://huggingface.co/Qwen/Qwen2-7B-Instruct

C.1 Scores for Each Model

The scores for each model are shown in Table 23. Claude 3.5 Sonnet achieved the highest score across all evaluation methods. Additionally, comparing Llama-3-70B with Llama-3-8B, and Qwen2-72B with Qwen2-7B, we can see a trend that larger models tend to achieve higher scores.

C.2 Comparison of Embedding Models

In addition to simcse-ja-bert-base-clcmlp, we also conduct evaluations using the following embedding models and calculate their correlation with human evaluation:

- OpenAI text-embedding-3-large[20]
- pkshatech/simcse-ja-bert-base-clcmlp
- pkshatech/GLuCoSE-base-ja[21]
- pkshatech/GLuCoSE-base-ja-v2
- cl-nagoya/sup-simcse-ja-large[22]
- cl-nagoya/ruri-large[23]

The Pearson correlation between each embedding model and human evaluation is shown in Table 24. simcse-ja-bert-base-clcmlp showed the highest correlation.

C.3 Relationship Between Number of Stories and Correlation with Human Evaluation

Figure 1 shows the relationship between the number of original stories and the Pearson correlation between embedding model evaluation and human evaluation for each model's scores. It becomes apparent that model scores from embedding model evaluation become reliable with approximately 20 stories.

[20] https://platform.openai.com/docs/models#embeddings

[21] https://huggingface.co/pkshatech/GLuCoSE-base-ja

[22] https://huggingface.co/cl-nagoya/sup-simcse-ja-large

[23] https://huggingface.co/cl-nagoya/ruri-large

We will display the original fairy tale and 11 modern versions of the story. Please rank the 11 modern versions in order of creativity. Enter your answer as single-byte numbers separated by single-byte spaces, with the more creative stories on the left.

Original Story
{Original Story}

Modern Version 1
{Modern Version 1}

Modern Version 2
{Modern Version 2}

(continued)

Table 21: Evaluation instructions for crowdworkers in SAT.

Please rate the creativity of the modern version of the story based on the original story on a scale of 1, 2, 3, 4, 5, and output only the number.

Rating Criteria
- 1: Not creative at all
- 2: Slightly creative
- 3: Creative
- 4: Very creative
- 5: Extremely creative

Original Story
{Original Story}

Modern Version
{Modern Version}

Table 22: Evaluation prompt for GPT-4o in SAT.

	Score by simcse-ja-bert-base-clcmlp	**Score by Human**	**Score by GPT-4o**
GPT-4o	0.513	0.559	0.692
GPT-4 Turbo	0.510	0.504	0.729
GPT-3.5 Turbo	0.405	0.456	0.630
Claude 3.5 Sonnet	**0.593**	**0.592**	**0.745**
Claude 3 Opus	0.514	0.505	0.667
Claude 3 Sonnet	0.570	0.523	0.664
Claude 3 Haiku	0.485	0.496	0.637
Llama-3-70B	0.496	0.478	0.630
Llama-3-8B	0.292	0.386	0.513
Qwen2-72B	0.478	0.501	0.694
Qwen2-7B	0.419	0.501	0.630

Table 23: SAT evaluation results for 11 models.

OpenAI text-embedding-3-large	0.863
pkshatech/simcse-ja-bert-base-clcmlp	**0.889**
pkshatech/GLuCoSE-base-ja	0.856
pkshatech/GLuCoSE-base-ja-v2	0.863
cl-nagoya/sup-simcse-ja-large	0.858
cl-nagoya/ruri-large	0.874

Table 24: Correlation between human evaluation and embedding models in SAT model evaluation. All p-values were below 0.05.

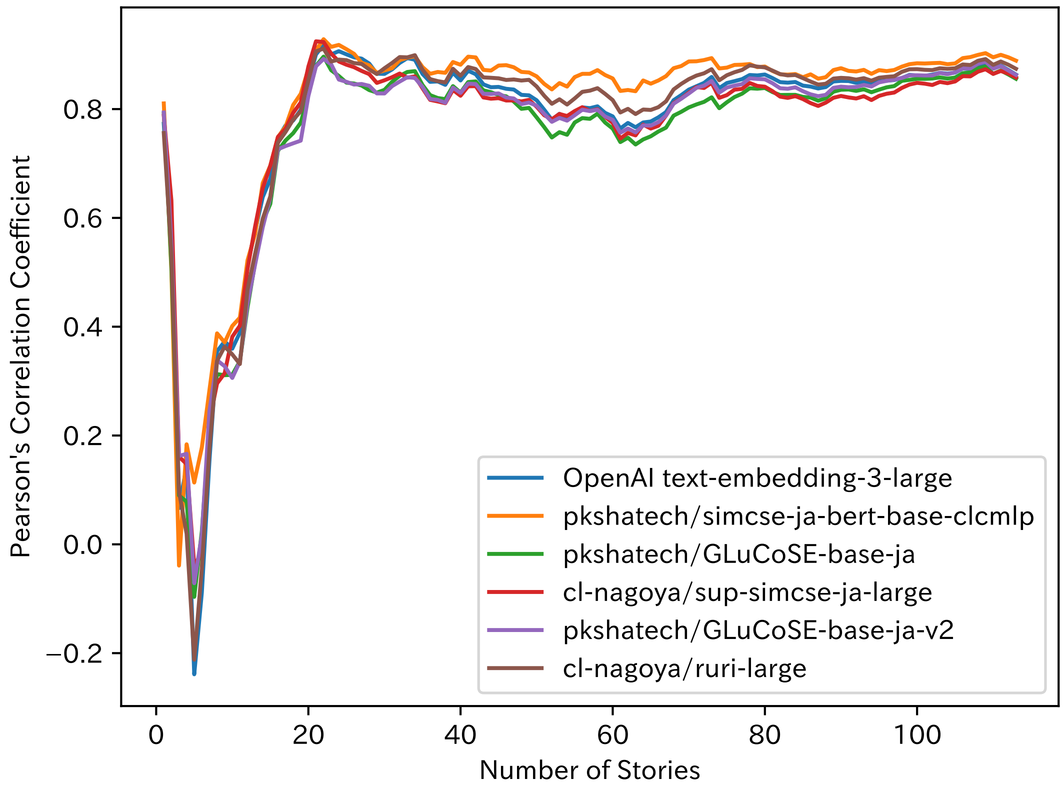

Figure 1: Relationship between number of stories and correlation with human evaluation in SAT.