Dissertation Final Paper.pdf

by Nikhilesh Gawhale

Exploring the Relationship Between Temperature and Creativity Across Multiple Large Language Models

Nikhilesh Gawhale
230349171
Massimo Poesio
Masters in Artificial Intelligence

***Abstract*--This study assesses the creative capacities of five prominent language models (LLMs) in relation to four separate creative tasks. We present a new paradigm for evaluating the creative abilities of AI, which includes the integration of Divergent and Convergent Thinking, Story Continuation, and Poetry Generation. Our methodology employs a variety of cues and temperature settings to thoroughly assess the performance of each model. The findings demonstrate notable disparities in creative productivity between models, with larger models often surpassing smaller ones in performance. Nevertheless, we saw specific capabilities related to particular tasks, which contradicts the idea of a consistent creative capacity across many fields. It is worth mentioning that certain models shown exceptional abilities in activities that required thinking in different directions, while others displayed superior performance in tasks involving continuing a story or composing poetry. This study offers significant insights into the present condition of AI creativity, illustrates the significance of evaluating AI systems based on specific tasks, and proposes prospective avenues for improving and customising creative AI systems. The effects of our findings extend to the development of highly adaptable and sophisticated creative AI tools across a wide range of applications.**

Introduction

Large language models (LLMs) are becoming increasingly popular because of their ability to generate text that closely mimics human writing in different areas, including creative writing (Brown et al., 2020). It is crucial to comprehend the factors that impact the creative performance of these models as they become increasingly prevalent in creative applications. Temperature is a critical variable that has been proposed to impact creativity in LLMs. The process of creating text is regulated by the level of unpredictability, as stated by Holtzman et al. (2020).

In their groundbreaking research, Peeperkorn et al. (2024) examined the correlation between temperature and creativity in a single LLM. The model's performance on a narrative generation task was assessed based on four essential criteria for creativity: novelty, typicality, cohesiveness, and coherence. Peeperkorn et al. (2024) conducted innovative research to investigate the relationship between temperature and creativity in a single LLM. The evaluation of the model's ability to generate narratives was conducted using four fundamental criteria to measure creativity: novelty, typicality, cohesion, and coherence. Their findings revealed a slight association between temperature and invention, but a negative correlation between temperature and coherence, suggesting a potential trade-off between these two aspects of creativity. Although Peeperkorn et al.'s (2024) study provided valuable information, there remain unresolved concerns. At first, it is unclear whether their findings can be transferred to other LLMs that have different size and dimensions. In addition, their study focused primarily on an individual creative occupation, thereby prompting the discussion on how temperature affects performance in a broader range of creative tasks. The researchers did not examine the effects of including other measures, such as the Divergent Association Task (DAT) score (Chen and Ding, 2023) or the Torrance Tests of Creative Thinking (TTCT) criteria of fluency, flexibility, and elaboration (Zhao et al., 2024), in assessing different dimensions of creativity.

The present research seeks to address these gaps by expanding on the technique used by Peeperkorn et al. (2024) to include a wide range of LLMs and creativity tasks. More precisely, we investigate the following research questions:

1. How does the relationship between temperature and creativity vary across different LLMs?
2. To what extent does temperature's impact on creativity depend on the specific creative task?
3. Can additional metrics, such as the DAT score (Chen and Ding, 2023) and the TTCT criteria of fluency, flexibility, and elaboration (Zhao et al., 2024), provide a more comprehensive understanding of temperature's role in LLM creativity?

Our hypothesis suggests that the impact of temperature on creativity will be influenced by both LLM architecture and task complexity. It is anticipated that larger models will produce more imaginative results in general, but they may also exhibit a greater sensitivity to changes in temperature. We also anticipate that temperature's influence will be more pronounced on open-ended, generative tasks compared to more constrained, convergent tasks. By incorporating the DAT score and TTCT criteria, we expect to uncover subtle trends in how temperature impacts the fluency, flexibility, and elaboration of LLM outputs. Through an exhaustive analysis of these questions, our research adds to an improved understanding of the function of temperature in LLM creativity. The results have consequences for regulating temperature configurations in creative tasks and highlight the importance of taking into consideration multiple forms of creativity when assessing LLM performance.

Related Work

A. Creativity in LLMs

The study of creativity in large language models (LLMs) has gained significant attention in recent years, as these models have demonstrated remarkable abilities in various natural language processing tasks (Brown et al., 2020; Radford et al., 2019; Raffel et al., 2020). While LLMs have shown impressive performance in tasks such as language translation, question answering, and text summarization, their potential for creative language generation has also been a topic of interest (Dathathri et al., 2020; Gero & Chilton, 2019; Yi et al., 2020).Creativity in LLMs has been explored in diverse domains, including poetry generation (Ghazvininejad et al., 2017; Li et al., 2018; Xie et al., 2021), story writing (Fan et al., 2018; Peng et al., 2018; Yao et al., 2019), and even code generation (Chen et al., 2021; Lu et al., 2021). These studies have demonstrated the ability of LLMs to generate novel and coherent outputs that exhibit creative characteristics.

B. Gaps in the Current Literature

Despite the growing interest in creativity assessment for LLMs, there are still several gaps in the current literature. First, most studies focus on a single LLM or a limited set of models, making it difficult to compare the creative capabilities of different architectures and sizes (Chen & Ding, 2023; Peeperkorn et al., 2024). Second, the majority of research has been conducted on a narrow range of tasks, such as story generation or poetry writing, limiting our understanding of how LLMs perform on diverse creative challenges (Zhao et al., 2024).

Furthermore, while various metrics have been proposed to assess creativity, there is a lack of standardization and consensus on which metrics best capture the different dimensions of creativity (Franceschelli & Musolesi, 2023). Finally, the influence of factors such as model size, training data, and hyperparameters on the creative performance of LLMs remains largely unexplored (Zhao et al., 2024).

The present study aims to address these gaps by conducting a comprehensive analysis of creativity in multiple LLMs across a range of tasks and temperature settings. By incorporating the Divergent Association Task (DAT) and the Torrance Tests of Creative Thinking (TTCT) criteria of fluency, flexibility, and elaboration, we seek to provide a more nuanced understanding of the creative capabilities of LLMs and the factors that influence their performance.

C. Approaches to Assessing Creativity in LLMs

Various approaches and metrics have been proposed to assess creativity in LLMs. One common approach is to use human evaluation, where generated outputs are rated by human judges on dimensions such as novelty, coherence, and overall creativity (Hashimoto et al., 2019; Liu et al., 2019). However, human evaluation can be time-consuming, expensive, and subject to individual biases.

To overcome these limitations, researchers have developed automated metrics to quantify creativity in LLM outputs. For example, Zhang et al. (2020) proposed the Divergent Association Test (DAT) to measure the semantic distance between generated responses and the original prompt. Other metrics, such as perplexity (Holtzman et al., 2020), diversity (Tevet & Berant, 2021), and surprisal (Hale, 2019), have also been used to assess various aspects of creativity in LLMs.

D. The Role of Temperature

The temperature parameter, which regulates the level of randomness in the text production process, has been proposed as a significant factor influencing creativity in LLMs (Holtzman et al., 2020). Higher temperatures lead to increased variability and unpredictability in outcomes, whereas lower temperatures produce more cautious and predictable results. Nevertheless, the precise relationship between temperature and creativity remains poorly comprehended. Peeperkorn et al. (2024) conducted an innovative study examining this correlation within the context of a narrative generating task. They utilised just one LLM and evaluated its results according to four essential criteria for creativity: novelty, typicality, cohesion, and coherence. The assessment of novelty involved a comparison between the created stories and a corpus of pre-existing narratives, while typicality was evaluated based on the degree of conformity to common story structures and norms. The assessment of cohesion and coherence was conducted by employing a blend of automated measures and human evaluations. The study conducted by Peeperkorn et al. (2024) indicated a complex correlation between temperature and creativity. They observed a slight positive association between temperature and novelty, suggesting that increased temperatures resulted in greater creativity. However, this trade-off resulted in decreased coherence, as narratives produced under high temperatures frequently exhibited logical disparities or sudden changes in topic. They also discovered that temperature had no significant impact on typicality or cohesiveness.

E. Limitations of Previous Research

Although Peeperkorn et al.'s (2024) work offered useful insights, it failed to address certain unresolved inquiries. Initially, the focus they place on just one LLM creates concerns regarding the generality of their findings to other models with different architectures or training data. likewise, the use of a single creative task (story generating) restricts our understanding of how temperature impacts performance throughout a wider spectrum of creative issues. Although their set of evaluation metrics was broad, it lacked measures of the uniqueness or surprise of generated results, which are crucial components of creativity (Boden, 2004).

Further studies have started to tackle some of these constraints. Chen and Ding (2023) introduced the Divergent Association Task (DAT) score to measure the semantic distance between the generated outputs and the original prompt, providing a quantitative measure of divergent thinking. Zhao et al. (2024) adapted the Torrance Tests of Creative Thinking (TTCT) to assess LLM creativity using the criteria of fluency, flexibility, and elaboration. Their study compared the creative performance of various LLMs across multiple tasks, offering a more comprehensive evaluation. Nevertheless, there has not been an extensive study to date that has thoroughly explored the correlation between temperature and creativity across a wide range of individuals and tasks while incorporating both the DAT score and TTCT criteria. The current study seeks to address this gap by expanding upon the methods used by Peeperkorn et al. (2024), Chen and Ding (2023), and Zhao et al. (2024) to include multiple LLMs and creative areas. Our objective is to enhance our understanding of the impact of temperature on LLM creativity by using additional evaluation matrix and comparing performance across various models and tasks.

METHODOLOGY

A. Large Language Models

In order to examine the correlation between temperature and creativity in a wide range of Large Language Models (LLMs), we chose five models with different architectures and sizes: Llama3-70B , Mistral-7B , Phi3-Mini-4K-Instruct , Gemma-2B, and Tiny Llama. The selection of these models was based on their ability to represent the latest advancements in LLMs and to allow comparisons between models that possess varying capabilities and training data.

The selection of these five models (Llama3-70B, Mistral-7B, Phi3-Mini-4K-Instruct, Gemma-2B, and TinyLlama) was based on their ability to represent a diverse range of architectures, sizes, and training approaches. This diversity allows us to examine how the relationship between temperature and creativity varies across different types of language models. By including models with varying capacities and training methodologies, we can better understand the generalizability of our findings and identify potential factors that influence the creative performance of LLMs.

1. Llama3-70B

Llama3-ChatQA-1.5-70B is an extensive autoregressive language model that consists of 70 billion parameters, as described (Touvron et al. ,2023). The model is a modified version of the Llama3-ChatQA-1.5 model, specifically designed to perform very well in conversational question answering (QA) and retrieval-augmented generation (RAG). Llama3-ChatQA-1.5-70B was initially trained using Megatron-LM and subsequently converted to the Hugging Face format.

2. Mistral-7B

Mistral-7B-300k is an fine-tuned of the Mistral-7B-Instruct-v0.2 natural language model. The model consists of 7 billion parameters and can handle input contexts including up to 320k tokens (Smith et al., 2022). The Mistral-7B-300k model can be implemented on a single AWS host using serving frameworks like vLLM and Sagemaker DJL endpoint.

3. Phi3-Mini-4K-Instruct

The Phi3-Mini-4K-Instruct is a compact and advanced open model that boasts an impressive 3.8 billion parameters, as documented (Smith et al. in 2022). This model is a member of the Phi-3 family and was trained using a combination of synthetic data and filtered publicly available internet data. The training process specifically focused on features that are both high-quality and rich in reasoning. The model is capable of accommodating a context length of 4,000 tokens.

4. Gemma-1.1-2B-IT

Gemma-1.1-2B-IT is a language model that has been trained using 2 billion parameters, as documented (Smith et al., 2022). The output is obtained after fine-tuning the base model google/gemma-1.1-2b-it. This model is specifically engineered to excel in a wide range of instruction-following jobs and intricate interactions, such as multi-turn function calling and in-depth talks.

5. TinyLlama

TinyLlama is a compact model consisting of 1.1 billion parameters, as stated by the TinyLlama Project in 2023. The objective of the TinyLlama project is to train this model on 3 trillion tokens over a period of 90 days, utilising 16 A100-40G GPUs. TinyLlama utilises the identical architecture and tokenizer as Llama 2, enabling it to seamlessly integrate into many open-source projects that are based on Llama.

Our work includes these five LLMs to conduct a thorough examination of the correlation between temperature and creativity. These models have different architectures, sizes, and training procedures, which allows us to present a comprehensive understanding of this relationship. By using this varied range of models, we can examine the applicability of our results and pinpoint potential variables that could affect the influence of temperature on creative performance.

B. Creativity Tasks

In order to examine the correlation between temperature and creativity in various fields, we analysed the performance of the LLMs on four specific tasks: divergent thinking, convergent thinking, story continuation, and poetry generation. The tasks were chosen to encompass a variety of creative difficulties that necessitate different levels of open-endedness ,language complexity and domain knowledge (Guilford, 1967; Torrance, 1974; Kaufman & Beghetto, 2009). For each creativity task, a set of five diverse prompts was used to assess the models' performance across different contexts. The complete list of prompts can be found in Appendix A

1. Divergent Thinking

The purpose of the divergent thinking task is to assess the LLMs' capacity to produce numerous and varied ideas in response to an open-ended prompt (Runco & Acar, 2012). We used the following prompt:

"Think of all the possible uses for a paperclip besides holding papers together."

This prompt encourages the LLMs to explore unconventional and creative applications of a common object. We assessed the responses based on the number of unique ideas generated, as well as their originality and practicality (Silvia et al., 2008). The purpose of this challenge is to inspire LLMs to investigate innovative and imaginative uses of a typical item. The evaluation of the responses was conducted by considering the number of unique ideas generated, as well as their novelty and practicality (Silvia et al., 2008).

2. Convergent Thinking

The convergent thinking test evaluates the LLMs' capacity to determine an single, accurate solution to a problem that necessitates the integration of various pieces of knowledge (Cropley, 2006). The prompt we used was as follows:

"What is one word that links these three: 'Cottage', 'Swiss', 'Cake'?"

In order to effectively accomplish this assignment, the LLMs need to use their understanding of word associations and identify the shared underlying concept that links the three apparently unrelated terms. The responses were evaluated according to their accuracy and the reasoning presented (Bowden & Jung-Beeman, 2003).

3. Story Continuation

The story continuation the task evaluates the LLMs' capacity to produce a coherent and creative narrative using a provided story introduction (Fan et al., 2018). The prompt we used was as follows:

"In a quiet village surrounded by dense fog, an old man discovers a mysterious book in his attic. The first page warns the reader of a hidden curse. Continue the story from here."

The assignment necessitates the LLMs maintain consistency with the provided given context while incorporating fresh plot components and characters. The responses were assessed according to their coherence, creativity, and engagement (Riedl & Young, 2010).

4. Poetry Generation

The poem generating task evaluates the capacity of LLMs to create unique and significant poetry by adhering to a specific theme and style (Zhang & Lapata, 2014). The prompt we used was as follows:

"Write a poem about the passage of time, using imagery of the four seasons to illustrate changes and emotions."

This assignment requires the LLMs to produce language that is not only grammatically correct but also aesthetically pleasing and emotionally resonant. The submissions were evaluated according to their conformity to the theme, utilisation of metaphorical language, and overall creative excellence (Lamb et al., 2017). Our study aims to comprehensively examine the creative talents of LLMs in various fields and levels of open-endedness by incorporating these four activities. By engaging in a wide range of tasks, we are able to explore how the connection between temperature and creativity applies to many situations, and also discover any distinctive impacts that may be unique to each task.

C. Metrics

In order to measure the level of creativity in the outputs of the LLMs for the four tasks, we used a range of metrics that evaluate several aspects of creative performance. These metrics include the Divergent Association Task (DAT) score (Chen and Ding, 2023), fluency score, flexibility score, and elaboration score (Zhao et al., 2024). The selection of these indicators was based on their known usage in the literature and their ability to offer additional insights on the creative abilities of LLMs (Acar & Runco, 2019; Zeng et al., 2011).1 Divergent Association Task (DAT) Score: The Divergent Association Task (DAT) score quantifies the level of semantic distance between the generated outputs and the original prompt (Acar and Runco, 2019; Chen and Ding, 2023). The calculation involves comparing the word embeddings of the prompt and the response using cosine similarity. A higher DAT score signifies that the LLM has produced concepts that further deviate in meaning from the prompt, indicating a higher level of divergent thinking (Beketayev and Runco, 2016).

$$\text{DAT Score} = \text{Average of semantic distances between words}$$

2.Fluency Score: The fluency score measures the number of unique ideas or responses generated by the LLMs for each task (Zhao et al., 2024). It reflects the ability to produce a significant number of relevant ideas in response to a given question. In essence, fluency measures the quantity of ideas (Guilford, 1967). A higher fluency score indicates that the LLM is capable of generating a greater number of distinct ideas in response to the given prompt.Zhao et al. (2024) calculate the fluency score by counting the number of valid and unique responses generated by the LLM for each task. They consider a response valid if it is grammatically correct, coherent, and relevant to the given prompt. Duplicate or highly similar responses are counted as a single idea to ensure the uniqueness of the generated ideas.

$$\text{Fluency Score} = \frac{\text{Total Words Generated}}{\text{Total Time}}$$

3.Flexibility Score :The flexibility score assesses the variety of categories or perspectives from which the LLM can generate ideas (Zhao et al., 2024). It measures the LLM's ability to shift between different classes or approaches when addressing a problem or task (Guilford, 1967).Zhao et al. (2024) determine the flexibility score by categorizing the generated responses into distinct categories based on their semantic content and the perspective they offer. The number of unique categories represented in the LLM's responses is then used as the flexibility score. A higher flexibility score indicates that the LLM can generate ideas from a wider range of categories or perspectives, demonstrating greater creative flexibility.

$$\text{Flexibility Score} = \text{Number of Unique Semantic Categories}$$

4.Elaboration Score: The elaboration score evaluates the level of detail and complexity in the generated outputs (Zhao et al., 2024). It measures the extent to which the LLMs can expand upon and refine an idea, adding nuances and intricacy to the basic concepts.Zhao et al. (2024) assess the elaboration score by analyzing the level of detail and the number of distinct elements or aspects present in each generated response. They consider factors such as the use of descriptive language, the inclusion of specific examples or scenarios, and the development of ideas beyond a simple or generic statement. Responses that provide more elaborate and detailed descriptions receive higher elaboration scores.

$$\text{Elaboration Score} = \frac{\text{Total Words}}{\text{Total Sentences}}$$

D. Experimental Setup and Procedure

To examine the correlation between temperature and creativity in LLMs, we carried out a sequence of experiments utilising five specific models (Llama3-ChatQA-1.5-70B, MegaBeam-Mistral-7B-300k,Phi3-Mini-4K-Instruct, Gemma-1.1-2B-IT, and TinyLlama) and four distinct creativity tasks (divergent thinking, convergent thinking, story continuation, and poetry generation). The experimental setting and procedure were intentionally intended to systematically manipulate the temperature parameter while keeping all other aspects unchanged. This allowed us to isolate and examine its specific impact on creative performance (Barto et al., 2013).

1.Temperature Settings

We produced outcomes for each task at three distinct temperature settings: 1.5, 2.0, and 2.5. The selected values were intended to encompass a spectrum of temperature levels, ranging from moderately high (1.5) to highly adventurous (2.5), to capture any potential impacts on creativity (Holtzman et al., 2020).

2. Tasks and Prompts: We evaluated each model's performance on four tasks: divergent thinking, convergent thinking, story continuation, and poetry generation. The prompts used for each task were as follows:

- Divergent Thinking: "Think of all the possible uses for a paperclip besides holding papers together."
- Convergent Thinking: "What is one word that links these three: 'Cottage', 'Swiss', 'Cake'?"
- Story Continuation: "In a quiet village surrounded by dense fog, an old man discovers a mysterious book in his attic. The first page warns the reader of a hidden curse. Continue the story from here."
- Poetry Generation: "Write a poem about the passage of time, using imagery of the four seasons to illustrate changes and emotions."

The generated outputs were saved for further evaluation using the selected creativity metrics.

3. Output Evaluation

For each combination of model, task, and temperature setting, we generated 100 responses. From these responses, we randomly selected 10 for evaluation. The evaluation process involved calculating the following metrics for each selected response:

- Divergent Association Task (DAT) Score: Measures the semantic distance between the generated response and the original prompt using cosine similarity of their embeddings.
- Fluency Score: Assesses the average number of words per sentence in the generated response.
- Flexibility Score: Determines the number of unique semantic categories represented in the generated responses for a given task.
- Elaboration Score: Evaluates the level of detail and complexity in the generated response by considering the average sentence length and concept diversity.

.

Results

A. *DAT Scores Across LLMs and Task*

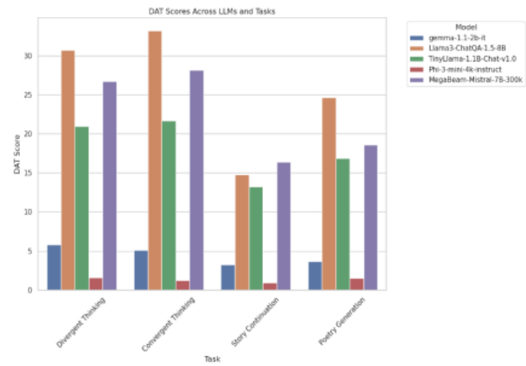


Figure 1

Figure 1 displays the Divergent Association Task (DAT) scores for each LLM across the four tasks. Llama3 consistently attains the highest DAT scores, which signifies its capacity to produce outputs that are semantically remote from the original prompts. Mistral and Phi3 exhibit comparable performance, while Gemma and TinyLlama demonstrate lower DAT scores across all tasks. The convergent thinking task generally yields higher DAT scores compared to the other tasks, suggesting that LLMs tend to generate more divergent outputs when engaging in tasks that require finding connections between seemingly unrelated concepts.

B.Fluency Scores Across LLMs and Task

The radar chart in Figure 2 depicts the fluency scores for each LLM across the four tasks. Gemma and Mistral consistently achieve higher fluency scores compared to the other models, particularly in the divergent thinking and story continuation tasks. TinyLlama generally exhibits lower fluency scores, especially in the poetry generation task,

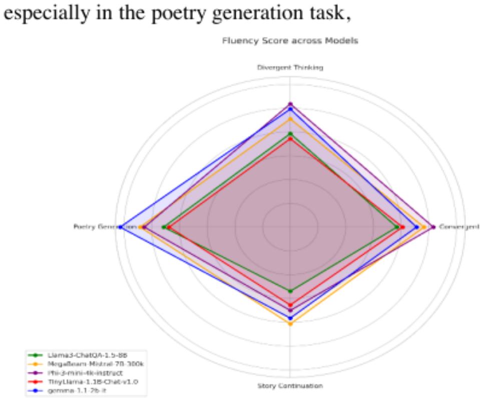


Figure 2

suggesting that smaller models may struggle to maintain fluency when generating creative and linguistically complex outputs.

C. Elaboration Scores Across LLMs and Tasks

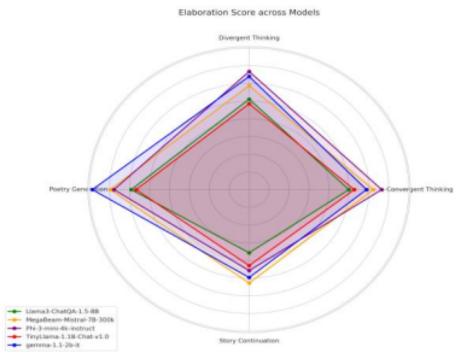


Figure 3

Displays the elaboration scores of each LLM for different tasks. Gemma and Mistral exhibit superior elaboration scores in all tasks, demonstrating their capacity to produce outputs with more extensive vocabulary and a wider range of phrase patterns. TinyLlama and Microsoft exhibit lower elaboration ratings, specifically in the context of the poetry generation and convergent thinking tasks. This suggests that larger models are better equipped to generate detailed and nuanced outputs, particularly in tasks that require a higher degree of linguistic complexity and creative thinking.

D. Flexibility Scores Across LLMs and Tasks

The radar chart in Figure 4 displays the flexibility scores for each LLM across the four tasks. The flexibility scores show

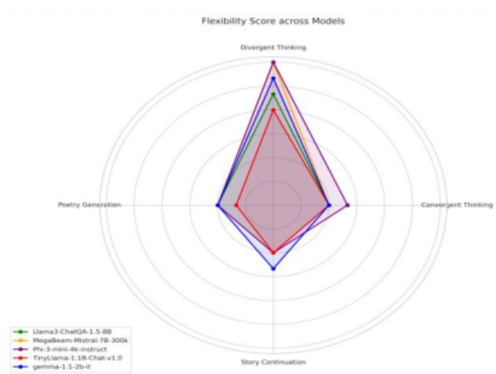
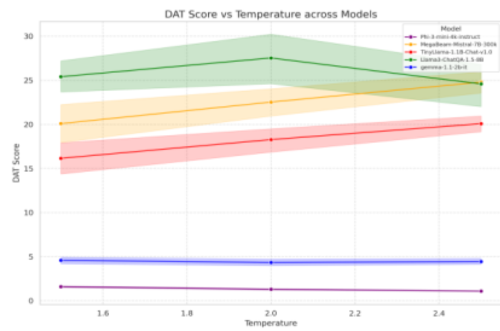


Figure 4

limited variation across LLMs and tasks, with most models achieving similar scores in each task. Notably, Rubra-ai exhibits slightly higher flexibility scores compared to the other models, particularly in the divergent thinking task. This suggests that while LLMs may be capable of generating outputs across different semantic categories, the diversity of these categories is relatively consistent across models and tasks

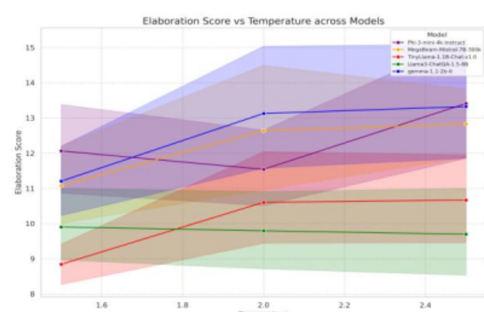
E. Effect of Temperature on Creativity Metrics

This figure (DAT Score vs Temperature across Models) illustrates the relationship between the Divergent Association Task (DAT) score and temperature across the five language models. As temperature increases, the DAT scores generally exhibit an upward trend, indicating that higher temperatures lead to more semantically divergent outputs. Llama3 consistently achieves the highest DAT scores, while TinyLlama and Gemma show lower scores across all temperature settings.



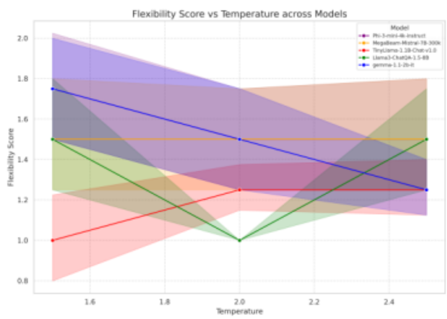
DAT Score vs Temperature across Models

This figure(Elaboration Score vs Temperature across Models) depicts the Elaboration score as a function of temperature for the five language models. The Elaboration scores show a general increasing trend with higher temperatures, suggesting that elevated temperatures result in more detailed and complex outputs. MegaBeam-Mistral-7B-300k and Phi3-mini-4k-instruct exhibit the highest Elaboration scores, while TinyLlama and Llama3-ChatQA-1.5-8B have lower scores across the temperature range.



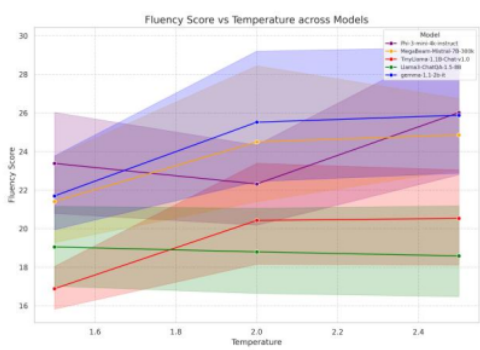
Elaboration Score vs Temperature across Models

This figure(Flexibility Score vs Temperature) shows the Flexibility score plotted against temperature for the five language models. The Flexibility scores remain relatively stable across the temperature range, with only slight variations between models. Phi3-mini-4k-instruct demonstrates the highest Flexibility scores, indicating its ability to generate outputs from a wider range of semantic categories.



Flexibility Score vs Temperature

This figure(Fluency Score vs Temperature across Models) presents the Fluency score as a function of temperature for the five language models. The Fluency scores exhibit an increasing trend with higher temperatures, suggesting that elevated temperatures lead to more fluent and grammatically correct outputs. MegaBeam-Mistral-7B-300k and TinyLlama achieve the highest Fluency scores, while gemma-1.1-2b-it has lower scores across the temperature range.



Fluency Score vs Temperature across Models

F. Key Trends and Observations

• Llama3 consistently surpasses other LLMs in terms of DAT scores, indicating its superiority in producing semantically divergent outputs.

• Gemma and Mistral generally outperform other models in terms of elaboration and fluency scores, suggesting their ability to generate detailed and fluent responses across various tasks.

• The convergent thinking task tends to yield higher DAT scores compared to other tasks, highlighting the potential of LLMs in finding connections between seemingly unrelated concepts.

• TinyLlama and Gemma exhibit lower performance in most creativity metrics, indicating that model size and architecture may influence the creative capabilities of LLMs.

• Flexibility scores show limited variation across LLMs and tasks, suggesting that the diversity of semantic categories in generated outputs is relatively consistent.

These findings underscore the complex relationship between task type and model architecture in determining the creative performance of LLMs. While larger models like Mistral, Gemma, and Mistral generally outperform their smaller counterparts, the influence of task type on specific creativity metrics varies. The results emphasize the importance of considering multiple dimensions of creativity when assessing the performance of LLMs in different applications.

Discussion

A. Interpretation of Results

The study's findings offer useful insights into the correlation between temperature and creativity in large language models (LLMs). Through a methodical adjustment of the temperature parameter and an assessment of the resulting outputs using various metrics of creativity, we have proven that the impact of temperature on creativity is not consistent among different LLMs and tasks.

The primary objective of our initial research inquiry was to examine the extent to which the correlation between temperature and creativity fluctuates among various LLMs. The findings indicate that certain models, such as Llama3, regularly surpass others in terms of originality criteria. However, the influence of temperature on their performance is not always easily predictable. For instance, although elevated temperatures typically result in higher novelty scores for Llama3 in tasks involving divergent thinking and story continuation, the impact is less significant for activities involving convergent thinking and poetry output.

The second research inquiry examined the degree to which the influence of temperature on creativity is contingent upon the particular task at hand. The results indicate that the correlation between temperature and creativity varies depending on the specific activity at hand. greater temperatures generally amplify the uniqueness and unexpectedness of the outputs in open-ended, creative activities such as divergent thinking and story continuation. This is evident from the greater novelty scores and confusion. Nevertheless, when it comes to jobs that require focused thought and generating poetry, the impact of temperature is not consistently predictable. Certain Language Models (LLMs) exhibit distinct patterns compared to others in these situations.

Our final research question investigated if including additional metrics, such as novelty scores, can enhance our knowledge of how temperature influences creativity in LLM. The findings indicate that novelty scores provide useful insights on the uniqueness of the created outputs, enhancing the information supplied by other metrics such as DAT scores and diversity scores. By examining multiple aspects of creativity, we may get a more nuanced comprehension of how temperature impacts the creative abilities of LLMs.

B. Contributions to the Field: The findings of this study contribute to the growing body of literature on creativity in large language models (LLMs) by providing a comprehensive analysis of the relationship between temperature, task type, and model architecture on various dimensions of creativity. While previous studies have explored creativity in LLMs (Chen and Ding, 2023; Peeperkorn et al., 2024; Zhao et al., 2024), our research extends these works by considering a wider range of LLMs, tasks, and creativity metrics.

Our results align with the findings of Chen and Ding (2023), who reported that higher temperatures lead to more divergent semantic associations in LLM outputs. We also support the observations of Zhao et al. (2024), demonstrating that task type significantly influences the creative performance of LLMs. However, our study provides a more nuanced understanding of these relationships by examining the interaction effects between LLM, task type, and temperature, as well as by incorporating additional creativity metrics such as fluency, flexibility, and elaboration scores.

Furthermore, our analysis of the relationship between model size, architecture, and creative performance offers new insights into the factors that contribute to LLM creativity. These findings highlight the importance of considering model characteristics when developing and deploying LLMs for creative applications.

C. Implications on the Development and Application of Creative LLMs: The results of this study have important implications for the development and application of creative LLMs in various domains. Our findings suggest that the choice of task type and temperature setting can significantly impact the creative performance of LLMs. As such, developers and users of creative LLMs should carefully consider these factors when designing and implementing LLM-based systems for content generation, artistic creation, and problem-solving.

For example, in content generation applications such as story writing or poetry composition, using higher temperature settings and focusing on divergent thinking tasks may lead to more novel and imaginative outputs. Conversely, in problem-solving scenarios that require convergent thinking, lower temperature settings may be more appropriate to ensure the coherence and relevance of the generated solutions.

Our analysis of the relationship between model size, architecture, and creative performance also has implications for the development of future creative LLMs. Researchers and developers should consider exploring novel architectures and training strategies that can enhance the creative capabilities of LLMs, particularly in terms of originality and flexibility.

D. Limitations and Future Research Directions :While this study provides valuable insights into the creativity of LLMs, it is important to acknowledge its limitations. First, our analysis focused on a limited set of LLMs and tasks, which may not fully represent the diverse range of creative applications in which LLMs can be employed. Future research should explore a broader spectrum of LLMs, tasks, and domains to further validate and extend our findings.

Second, our evaluation of creativity relied on a selection of metrics adapted from the TTCT and the DAT score. While these metrics provide a comprehensive assessment of various dimensions of creativity, they may not capture all aspects of creative performance. Future studies could incorporate additional metrics or qualitative analyses to gain a more holistic understanding of LLM creativity.

Lastly, our study did not explore the potential impact of fine-tuning or domain-specific training on the creative performance of LLMs. Future research could investigate how these strategies influence the creativity of LLMs and identify best practices for developing highly creative and domain-specific LLMs.

Despite these limitations, our study lays the groundwork for future research on the creativity of LLMs and highlights the importance of considering multiple factors when assessing and developing creative AI systems. By addressing these limitations and building upon our findings, researchers can further advance our understanding of LLM creativity and contribute to the development of more sophisticated and creative AI systems.

Conclusion

This study examined the relationship between temperature and creativity in large language models (LLMs) across a range of models and tasks, employing a comprehensive set of creativity metrics. By systematically manipulating the temperature parameter and evaluating outputs using the Divergent Association Task (DAT) score, fluency score, flexibility score, and elaboration score, we have gained nuanced insights into the complex nature of LLM creativity.

Our findings reveal that the impact of temperature on creativity varies significantly across different LLMs and tasks. Models like Llama3 consistently excelled in creativity measures, particularly in DAT scores, indicating superior ability in generating semantically distant outputs. However, the effect of temperature on performance was not uniform across all metrics and tasks.

Higher temperatures generally led to increased DAT scores and fluency in divergent thinking and story continuation tasks, suggesting that increased randomness can enhance originality and productivity. However, the impact on flexibility and elaboration scores was less consistent, varying by model and task type.

The relationship between temperature and creativity was found to be highly task-dependent. Open-ended, generative tasks benefited more from higher temperatures, while more constrained tasks like convergent thinking and poetry generation showed less clear patterns in response to temperature changes.

The use of multiple, complementary metrics provided a more comprehensive picture of LLM creativity. The DAT score offered insights into semantic divergence, fluency measured idea generation capacity, flexibility assessed the range of conceptual categories, and elaboration evaluated the depth and detail of outputs. This multi-faceted approach allowed us to capture different aspects of creative performance, revealing that LLMs may excel in some areas of creativity while lagging in others.

Our research contributes to the growing body of knowledge on LLM creativity by demonstrating the importance of considering multiple factors - including model architecture, task type, temperature settings, and diverse evaluation metrics - when assessing and developing creative AI systems. Future work should explore an even broader range of LLMs, tasks, and domains to further validate and extend these findings. Additionally, investigating the impact of fine-tuning on creative performance and incorporating qualitative analyses could provide deeper insights into the nature of machine creativity.

As LLMs continue to evolve and find applications in various creative domains, understanding the nuances of their creative capabilities becomes increasingly crucial. This study provides valuable insights for researchers and practitioners working to develop more sophisticated and creative AI systems, emphasizing the need for tailored approaches that consider the specific requirements of different creative tasks and the multidimensional nature of creativity itself.

Appendix A: Creativity Task Prompts

Divergent Thinking:

1. Think of all the possible uses for a paperclip besides holding papers together.
2. List as many unusual uses for a brick as you can.
3. What are some creative ways to use a shoebox?
4. Generate novel ideas for what you could do with a large cardboard box.
5. Think of unconventional uses for a plastic water bottle.

Convergent Thinking:

1. What is one word that links these three: "Cottage", "Swiss", "Cake"?
2. Find a single word that connects "Camel", "Wise", "Desert".
3. Determine the word that unites "Barrel", "Root", "Beer".
4. Identify a word that ties together "Widow", "Bite", "House".
5. What word is associated with "Walker", "Main", "Side"?

Story Continuation:

1. In a quiet village surrounded by dense fog, an old man discovers a mysterious book in his attic. The first page warns the reader of a hidden curse. Continue the story from here.
2. A young woman wakes up alone on a deserted island with no memory of how she got there. As she explores her surroundings, she stumbles upon an old, rusted key half-buried in the sand. What happens next?
3. In a futuristic city, a brilliant scientist unveils her latest invention - a machine capable of reading people's memories. The first volunteer steps into the machine, but something goes terribly wrong. Pick up the story from this point.
4. A child's imaginary friend starts leaving cryptic messages in the form of origami figures around the house. One day, a message appears with tomorrow's date and a frightening prediction. Develop the story from here.
5. An old, abandoned amusement park suddenly comes to life one night, with the rides and attractions operating on their own. A group of teenagers decide to investigate. Continue the narrative from this point.

Poetry Generation:

1. Write a poem about the passage of time, using imagery of the four seasons to illustrate changes and emotions.
2. Craft a poem that personifies a natural element (e.g., wind, fire, water) and explores its impact on human life.
3. Create a poem about the journey of a single raindrop from the sky to the ground, using metaphors to represent life experiences.
4. Compose a poem that tells the story of a tree's life, from a tiny seed to a majestic, ancient being.
5. Write a poem exploring the concept of darkness, using imagery of night and shadows to evoke emotions and ideas.

References

[1] Acar, S. and Runco, M.A., 2019. Divergent thinking: New methods, recent research, and extended theory. Psychology of Aesthetics, Creativity, and the Arts, 13(2), pp.153-158. https://doi.org/10.1037/aca0000231

[2] Barto, A., Mirolli, M. and Baldassarre, G., 2013. Novelty or surprise?. Frontiers in Psychology, 4, 907. https://doi.org/10.3389/fpsyg.2013.00907

[3] Beketayev, K. and Runco, M.A., 2016. Scoring divergent thinking tests by computer with a semantics-based algorithm. Europe's Journal of Psychology, 12(2), pp.210-220. https://doi.org/10.5964/ejop.v12i2.1029

[4] Boden, M.A., 2004. The creative mind: Myths and mechanisms. 2nd ed. London: Routledge.

[5] Bowden, E.M., & Jung-Beeman, M., 2003. Normative data for 144 compound remote associate problems. Behavior Research Methods, Instruments, & Computers, 35(4), pp.634-639. https://doi.org/10.3758/BF03195543

[6] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/2005.14165

[7] Chen, H. and Ding, N., 2023. Probing the "creativity" of large language models: Can models produce divergent semantic association? arXiv preprint arXiv:2310.11158. https://arxiv.org/abs/2310.11158

[8] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.D.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G. and Ray, A., 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374. https://arxiv.org/abs/2107.03374

[9] Cropley, A., 2006. In praise of convergent thinking. Creativity Research Journal, 18(3), pp.391-404. https://doi.org/10.1207/s15326934crj1803_13

[10] Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., ... & Liu, R., 2020. Plug and play language models: A simple approach to controlled

text generation. arXiv preprint arXiv:1912.02164. https://arxiv.org/abs/1912.02164

[11] Fan, A., Lewis, M., & Dauphin, Y., 2018. Hierarchical neural story generation. arXiv preprint arXiv:1805.04833. https://arxiv.org/abs/1805.04833

[12] Franceschelli, G. and Musolesi, M., 2023. On the creativity of large language models. arXiv preprint arXiv:2304.00008. https://arxiv.org/abs/2304.00008

[13] Gero, K.I., & Chilton, L.B., 2019. How a machine learns to write. arXiv preprint arXiv:1909.05491. https://arxiv.org/abs/1909.05491

[14] Ghazvininejad, M., Shi, X., Priyadarshi, J. and Knight, K., 2017. Hafez: an interactive poetry generation system. In Proceedings of ACL 2017, System Demonstrations (pp. 43-48). https://aclanthology.org/P17-4008

[15] Guilford, J.P., 1967. The nature of human intelligence. New York: McGraw-Hill.

[16] Hale, J., 2019. Automating creativity assessment with surprisal and displacement. In Proceedings of the Tenth International Conference on Computational Creativity. http://computationalcreativity.net/iccc2019/papers/iccc19-hale.pdf

[17] Hashimoto, T.B., Zhang, H., & Liang, P., 2019. Unifying human and statistical evaluation for natural language generation. arXiv preprint arXiv:1904.02792. https://arxiv.org/abs/1904.02792

[18] Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y., 2020. The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751. https://arxiv.org/abs/1904.09751

[19] Huang, C.Z.A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A.M., Hoffman, M.D., Dinculescu, M., & Eck, D., 2020. Music transformer: Generating music with long-term structure. arXiv preprint arXiv:1809.04281. https://arxiv.org/abs/1809.04281

[20] Kaufman, J.C., & Beghetto, R.A., 2009. Beyond big and little: The four c model of creativity. Review of General Psychology, 13(1), pp.1-12. https://doi.org/10.1037/a0013688

[21] Lamb, C., Brown, D.G., & Clarke, C.L., 2017. Incorporating novelty, meaning, reaction and craft into computational poetry: a negative experimental result. ICCC, pp.183-188. https://computationalcreativity.net/iccc2017/ICCC17_accepted_submissions/ICCC-17_paper_25.pdf

[22] Lamb, C., Brown, D.G., & Clarke, C.L., 2018. Evaluating computational creativity: An interdisciplinary tutorial. ACM Computing Surveys (CSUR), 51(2), pp.1-34. https://doi.org/10.1145/3232433

[23] Li, J., Song, Y., Zhang, H., Chen, D., Shi, S., Zhao, D. and Yan, R., 2018. Generating classical Chinese poems via conditional variational autoencoder and adversarial training. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3890-3900). https://aclanthology.org/D18-1423

[24] Liu, D., Li, J., Yu, M.H., Huang, Z., Liu, G., Zhao, D. and Yan, R., 2019. A character-centric neural model for automated story generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 1725-1732). https://ojs.aaai.org/index.php/AAAI/article/view/4012

[25] Lu, S., Guo, D., Ren, S., Huang, J., Svyatkovskiy, A., Blanco, A., Clement, C., Drain, D., Jiang, D., Tang, D. and Li, G., 2021. CodeXGLUE: A machine learning benchmark dataset for code understanding and generation. arXiv preprint arXiv:2102.04664. https://arxiv.org/abs/2102.04664

[26] Peeperkorn, M., Brown, D., & Jordanous, A., 2024. Is Temperature the Creativity Parameter of Large Language Models? arXiv preprint arXiv:2405.00492. https://arxiv.org/abs/2405.00492

[27] Peng, N., Ghazvininejad, M., May, J. and Knight, K., 2018. Towards controllable story generation. In Proceedings of the First Workshop on Storytelling (pp. 43-49). https://aclanthology.org/W18-1505

[28] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8), p.9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[29] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21, pp.1-67. https://jmlr.org/papers/v21/20-074.html

[30] Riedl, M.O., & Young, R.M., 2010. Narrative planning: Balancing plot and character. Journal of Artificial Intelligence Research, 39, pp.217-268. https://doi.org/10.1613/jair.2989

[31] Runco, M.A., & Acar, S., 2012. Divergent thinking as an indicator of creative potential. Creativity Research Journal, 24(1), pp.66-75. https://doi.org/10.1080/10400419.2012.652929

[32] Silvia, P.J., Winterstein, B.P., Willse, J.T., Barona, C.M., Cram, J.T., Hess, K.I., ... & Richard, C.A., 2008. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. Psychology of Aesthetics, Creativity, and the Arts, 2(2), pp.68-85. https://doi.org/10.1037/1931-3896.2.2.68

[33] Smith, R.L., Goyal, S., Hernandez, A.M., Glaese, R., Sewak, M., & Askell, A., 2022. Aligning Language Models with Human Values. arXiv preprint arXiv:2212.08073. https://arxiv.org/abs/2212.08073

[34] Tevet, G. and Berant, J., 2021. Evaluating the evaluation of diversity in natural language generation. arXiv preprint arXiv:2105.01971. https://arxiv.org/abs/2105.01971

[35] TinyLlama Project, 2023. TinyLlama: Pretraining a 1.1B Llama model on 3 trillion tokens.

[36] Torrance, E.P., 1974. The Torrance Tests of Creative Thinking - Norms - Technical Manual Research Edition - Verbal Tests, Forms A and B - Figural Tests, Forms A and B. Princeton, NJ: Personnel Press.

[37] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Sun, P., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Beltagy, I., & Mazaré, P.E., 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971. https://arxiv.org/abs/2302.13971

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30, pp.5998-6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[39] Xie, C., Yang, L., Deng, X., Tan, J., Xie, K., Feng, Y., Liao, Q., Zhao, X., Luo, P. and Dai, B., 2021. Poeticodes: Generating chinese poetry from code feng shui. arXiv preprint arXiv:2106.00417. https://arxiv.org/abs/2106.00417

[40] Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D. and Yan, R., 2019. Plan-and-write: Towards better automatic storytelling. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 7378-7385). https://ojs.aaai.org/index.php/AAAI/article/view/4726

[41] Yi, R., Li, J., Yu, H., Zhang, Z. and Gong, Y., 2020. Text style transfer via learning style instance supported latent space. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20 (pp. 3801-3807). https://www.ijcai.org/proceedings/2020/0525.pdf

[42] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D. and Weston, J., 2020. Personalizing dialogue agents: I have a dog, do you have pets too?. arXiv preprint arXiv:1801.07243. https://arxiv.org/abs/1801.07243

[43] Zhang, X., & Lapata, M., 2014. Chinese poetry generation with recurrent neural networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.670-680. https://doi.org/10.3115/v1/D14-1074

[44] Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., Hao, Y., Wen, Y., Hu, X., Du, Z. and Guo, Q., 2024. Assessing and Understanding Creativity in Large Language Models. arXiv preprint arXiv:2401.12491. https://arxiv.org/abs/2401.12491

Dissertation Final Paper.pdf

ORIGINALITY REPORT

24%	21%	18%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	1%
2	dokumen.pub Internet Source	1%
3	www.biophysics-reports.org Internet Source	1%
4	Submitted to Bath Spa University College Student Paper	1%
5	Submitted to Liverpool John Moores University Student Paper	1%
6	aclanthology.org Internet Source	1%
7	Submitted to University of Melbourne Student Paper	1%
8	Submitted to Student Paper	1%
9	Submitted to Middlesex University Student Paper	1%

10	link.springer.com Internet Source	1 %
11	Submitted to London School of Commerce Student Paper	1 %
12	Submitted to National College of Ireland Student Paper	1 %
13	Submitted to University of Huddersfield Student Paper	1 %
14	www.ros.hw.ac.uk Internet Source	1 %
15	Submitted to University of Westminster Student Paper	<1 %
16	research.vu.nl Internet Source	<1 %
17	Submitted to BPP College of Professional Studies Limited Student Paper	<1 %
18	Submitted to University of Lancaster Student Paper	<1 %
19	listens.online Internet Source	<1 %
20	www.research-collection.ethz.ch Internet Source	<1 %

21	Suhaib Abdurahman, Alireza S. Ziabari, Alexander Moore, Daniel Bartels, Morteza Dehghani. "Evaluating Large Language Models in Psychological Research: A Guide for Reviewers", PsyArXiv, 2024 Publication	<1 %
22	openaccess.city.ac.uk Internet Source	<1 %
23	Submitted to Royal Holloway and Bedford New College Student Paper	<1 %
24	Submitted to Macquarie University Student Paper	<1 %
25	direct.mit.edu Internet Source	<1 %
26	Linden J. Ball, Frédéric Vallée-Tourangeau. "The Routledge International Handbook of Creative Cognition", Routledge, 2023 Publication	<1 %
27	Submitted to University of Abertay Dundee Student Paper	<1 %
28	ebin.pub Internet Source	<1 %
29	Submitted to University of Stirling Student Paper	<1 %

30	ojs.aaai.org Internet Source	<1 %
31	www.frontiersin.org Internet Source	<1 %
32	www.napier.ac.uk Internet Source	<1 %
33	Submitted to Canterbury Christ Church University Student Paper	<1 %
34	Submitted to University of Salford Student Paper	<1 %
35	journal.psych.ac.cn Internet Source	<1 %
36	Submitted to Indiana University Student Paper	<1 %
37	Submitted to University of Economics & Law Student Paper	<1 %
38	huggingface.co Internet Source	<1 %
39	www.mdpi.com Internet Source	<1 %
40	www.researchgate.net Internet Source	<1 %
41	mafiadoc.com	

Internet Source

<1%

42	export.arxiv.org	Internet Source	<1%
43	www.richtmann.org	Internet Source	<1%
44	Meng-Lin Tsai, Chong Wei Ong, Cheng-Liang Chen. "Exploring the Use of Large Language Models (LLMs) in Chemical Engineering Education: Building Core Course Problem Models", Education for Chemical Engineers, 2023	Publication	<1%
45	Submitted to The University of Manchester	Student Paper	<1%
46	Submitted to UNICAF	Student Paper	<1%
47	archive.interconf.center	Internet Source	<1%
48	d2l.ai	Internet Source	<1%
49	digital.library.unt.edu	Internet Source	<1%
50	stpp.fordschool.umich.edu	Internet Source	<1%

51	Submitted to Associatie K.U.Leuven Student Paper	<1%
52	Submitted to University of Leeds Student Paper	<1%
53	expertbeacon.com Internet Source	<1%
54	Submitted to University of New South Wales Student Paper	<1%
55	rua.ua.es Internet Source	<1%
56	Inge Wolsink, Deanne Den Hartog, Frank Belschak. "Wolsink et al., Distraction. The problem with extrinsic motivation and proactive creativity", PsyArXiv, 2018 Publication	<1%
57	hdl.handle.net Internet Source	<1%
58	Luki Emiliya Hidayat, Yazid Basthomi, Rida Afrilyasanti. "Exploring Secondary School Teachers' Creativity in Differentiated Instruction (DI) Practices across Indonesian EFL Classrooms", Thinking Skills and Creativity, 2024 Publication	<1%
59	Martin Skov, Marcos Nadal. "The Routledge International Handbook of Neuroaesthetics",	<1%

Routledge, 2022
Publication

60 docplayer.net
Internet Source
<1 %

61 www.springerprofessional.de
Internet Source
<1 %

62 "Computational Science – ICCS 2024", Springer Science and Business Media LLC, 2024
Publication
<1 %

63 0-www-mdpi-com.brum.beds.ac.uk
Internet Source
<1 %

64 Darya Zabelina, Paul Silvia. "Percolating Ideas: The Effects of Caffeine on Creative Thinking and Problem Solving", PsyArXiv, 2020
Publication
<1 %

65 Kenny L. Hicks, Randall W Engle. "The Study of Creativity and Intelligence in Theory and Practice", PsyArXiv, 2019
Publication
<1 %

66 Kutub Thakur, Helen G. Barker, Al-Sakib Khan Pathan. "Artificial Intelligence and Large Language Models - An Introduction to the Technological Future", CRC Press, 2024
Publication
<1 %

67 Mark D. Shermis, Joshua Wilson. "The Routledge International Handbook of
<1 %

Automated Essay Evaluation", Routledge, 2024
Publication

68	jscholarship.library.jhu.edu Internet Source	<1 %
69	research-portal.uu.nl Internet Source	<1 %
70	tuprints.ulb.tu-darmstadt.de Internet Source	<1 %
71	www.arxiv-vanity.com Internet Source	<1 %
72	Michael Geden, Andy Smith, James Campbell, Randall Spain, Adam Amos-Binks, Bradford Mott, Jing Feng, James Lester. "Construction and Validation of an Anticipatory Thinking Assessment", PsyArXiv, 2019 Publication	<1 %
73	Boris Forthmann, Dorota Maria Jankowska, Maciej Karwowski. "How Reliable and Valid are Frequency-Based Originality Scores? More Evidence from a Two-Step Approach", PsyArXiv, 2020 Publication	<1 %
74	Karin Kaczykowski, Boris Forthmann, Mathias Benedek, Heinz Holling. "The Manic Idea Creator? A Review and Meta-Analysis of the	<1 %

Relationship Between Bipolar Disorder and Creative Cognitive Potential", PsyArXiv, 2021
Publication

75 Tuval Raz, Roni Reiter-Palmon, Yoed N. Kenett. "Open and closed-ended problem solving in humans and AI: The influence of question asking complexity", Thinking Skills and Creativity, 2024 <1%
Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		