AI Writers and Critics: An Exploratory Study on Creative Content Generation and Evaluation by Large Language Models

Shraddha Vijay Pawar*,†, Savita Bhat†, Ganesh Prasath and Shirish Karande

TCS Research

Abstract

Recently, large language models (LLMs) have demonstrated promising potential in creative writing tasks, including stories and poems. However, existing studies have often focused on limited tasks and have not fully explored the comprehensive capabilities of these models across a diverse range of creative writing forms. This paper presents a comprehensive analysis of the performance of 11 LLMs across various creative writing tasks, including poem writing, blog creation, ad copy generation, short movie scriptwriting, and news article production. Our research aims to evaluate whether larger models consistently produce superior content or if smaller models can achieve comparable results. We examine different prompting styles and multi-agent frameworks to understand their impact on output quality. Using a detailed rubric based on a defined set of criteria such as innovation, adherence, coherence, expressiveness, conformity, and diversity, we systematically assess the generated content. The study highlights the effectiveness of using multiple LLMs as evaluators to enhance the reliability of content evaluation. Our findings provide insights into the capabilities and limitations of LLMs in creative tasks, suggesting avenues for future research to improve their creative potential and evaluation methodologies.

Keywords

Large Language Models, Machine Creativity, Generative Artificial Intelligence, Creative Writing Tasks, AI-Generated Content, LLM-Based Evaluations, Content Quality Assessment

1. Introduction

Content creation spans various textual forms, each demanding a unique approach and skill set, from news articles and academic papers to blogs, songs, poetry, and storytelling [1] [2]. Poetry uses aesthetic and rhythmic language to convey deeper meanings, while story writing focuses on plot and character development. Songwriting blends lyrics with music for an auditory experience, and ad copy aims to persuade using rhetorical devices. Blogs serve personal expression or information sharing, scriptwriting crafts visual storytelling, dialogue writing hones conversational skills, and news writing informs with factual reporting [3] [4]. These diverse forms enrich our cultural landscape and support various professions, requiring a nuanced understanding of language, audience, and purpose. The evolution of AI and NLP technologies, particularly advancements in LLMs like GPT-3 and GPT-4, has significantly impacted content

CREAI 2024 - Workshop on Artificial Intelligence and Creativity

*Corresponding author.

†These authors contributed equally.

shraddhavijay.pawar@tcs.com (S. V. Pawar); savita.bhat@tcs.com (S. Bhat); ganesh.prasathr@tcs.com (G. Prasath); shirish.karande@tcs.com (S. Karande)

writing, enabling these models to generate coherent and contextually relevant text, making them powerful tools for diverse applications.

To understand the writing quality produced by LLMs, our research considers various writing tasks, each requiring specific skills. We develop criteria common to these tasks and use them to design comprehensive rubrics for each problem statement, to assess the generated content. By evaluating 11 different diverse models of various sizes, we explore both large and small models' capabilities. Our objective is to determine if larger models consistently produce superior quality content or if smaller models can achieve comparable results. Additionally, we examine different prompting styles and multi-agent frameworks to understand how prompt engineering affects the LLMs' output quality. This methodology systematically determines which models excel in particular tasks and how various prompting approaches influence content generation effectiveness.

2. Related work

LLMs have been increasingly utilized in creative writing tasks, with notable studies demonstrating their potential and limitations. For instance, [5] showcased the beneficial role of AI in translation and reviewing to enhance writing. However, their focus was limited to generating drafts based on user-provided plots in short fiction and non-fiction writing tasks using Chat GPT-3.5. This narrow scope underscores the need for broader exploration across various creative writing tasks, a gap that our research addresses. Similarly, [6] conducted a comparative study of GPT-3 with other state-of-the-art models like KGGPT2, HINT, PROGEN, and MTCL. Their findings highlighted GPT-3's superiority in story generation, particularly the text-davinci-001 variant. However, they acknowledged the advancements in newer models like GPT-4 and LLAMA 3, which their study did not cover. Our research builds on this by assessing these newer models to evaluate their enhanced content generation capabilities.

We considered the comprehensive evaluation by [7] of LLMs for English content writing, focusing on both open-source and closed-source models. Their study was confined to a single problem statement for narrative writing using zero-shot prompting. In our study, we use their human evaluation results as a benchmark and explore more complex prompting styles like chain of thought (CoT)[8] and Reason and Act (ReAct) [9] to enhance content generation. The examination of creativity in LLMs by[10] is relevant for its consideration of value, novelty, and surprise through Margaret Boden's creativity theories. They argued that truly creative processes in LLMs require attributes like motivation, thinking, and perception, which are currently lacking. This insight motivated us to explore advanced prompting styles to better evaluate LLMs' creative capabilities.

The study by [11] highlights the challenge of non-expert judges in assessing creativity, often favoring novice work over professional outputs. They examined metrics like Ritchie's model[12], Pease et al.'s criteria, Colton's creative tripod,[13] and the IDEA model[14]. We integrated these with Boden's criteria[15] to design effective rubrics for evaluating creativity in LLMs. The framework introduced by [16] for evaluating LLM creativity through verbal question-and-answer formats, using a modified Torrance Tests of Creative Thinking, offers a comprehensive approach. However, it emphasizes the need to assess creativity beyond verbal

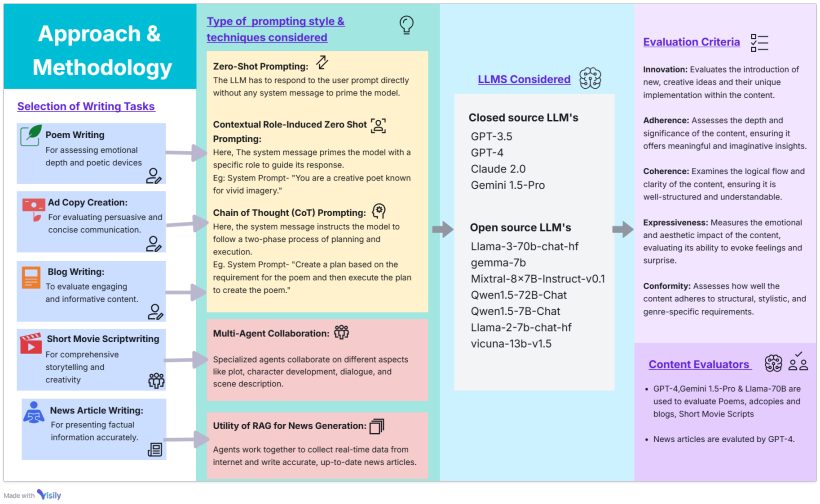


Figure 1: Overview of the Approach & Methodology

formats. Our study addresses this by exploring a broader range of writing tasks, including poem writing, ad copy creation, blog writing, short movie scriptwriting, and news articles. Finally, [17] introduced the TTCW framework to evaluate creativity in short fiction writing by both human authors and LLMs. Their findings showed that LLM-generated stories passed fewer TTCW tests compared to human-authored ones, highlighting a gap in creative proficiency. This insight led us to extend our experimentation with LLMs to assess creativity in various other writing tasks beyond story writing.

3. Approach and Methodology

Our methodology evaluates LLM writing quality across various tasks using common criteria. We assess 11 models of different sizes and explore different prompting styles and techniques This process is done to identify the best models for specific tasks and the impact of prompt engineering on content quality. Refer to the figure 1 for the details of the process.

3.1. Selection of Writing Tasks

We selected tasks that highlight the unique capabilities of LLMs in content generation, each designed to assess specific aspects of the model's creative and analytical skills. Poem Writing involves problem statements that require the model to evoke particular emotions such as melancholy, resilience, and playfulness, using tones like reflective, whimsical, or serious. Ad Copy Creation challenges the model to craft persuasive and concise messages, demanding a precise balance of action-oriented language and emotional appeal. Blog Writing requires the model to sustain reader engagement over varied lengths, blending informative and narrative styles, while maintaining a consistent tone that can range from analytical to conversational. Short Movie Scriptwriting emphasizes the model's ability to create compelling narratives, requiring

effective representation of character development, dialogue authenticity, and thematic depth. Lastly, News Article Writing focuses on the model's capacity to present factual information with clarity and neutrality, analyzing its ability to convey events accurately while maintaining an objective and accessible tone.

3.2. Criteria for Content Assessment

To comprehensively assess creativity in language models across diverse writing tasks, we utilize several established theoretical models, including Ritchie's Model [12], Pease et al.'s criteria, Boden's criteria [15], the IDEA Model by Isaksen and Dorval, and Colton's Creative Tripod [13]. By studying these models, we developed a customized set of criteria given below, applicable across the writing skills chosen.

- **Innovation:** Evaluates the introduction of new, creative ideas and their unique implementation within the content. Derived from Ritchie's Model, Pease et al.'s Novelty, and Boden's Novelty, emphasizing the novelty, uniqueness, and unexpected insights in content.
- **Adherence:** Assesses the depth and significance of the content, ensuring it offers meaningful and imaginative insights. Inspired by Boden's Value and Colton's Imagination, focusing on the usefulness, depth, and imaginative quality of the content.
- **Coherence:** Examines the logical flow and clarity of the content, ensuring it is well-structured and understandable. Based on the IDEA Model's Evaluation and Colton's Skill, integrating refinement of ideas and technical proficiency.
- **Expressiveness:** Measures the emotional and aesthetic impact of the content, evaluating its ability to evoke feelings and surprise. Rooted in Colton's Appreciation and Boden's Surprise, focusing on emotional connection and unexpected elements.
- **Conformity:** Assesses how well the content adheres to structural, stylistic, and genre-specific requirements. Derived from the IDEA Model's Intention and Action, ensuring alignment with genre-specific conventions and effective implementation of ideas.
- **Diversity:** Evaluates the range and variation in outputs for the same problem statement, ensuring distinct and novel solutions. Inspired by Boden's Novelty and Surprise and Ritchie's Originality, focusing on generating varied and original content.

We use these criteria as a high-level foundation to establish a hierarchical structure for content evaluation, meticulously developing task-specific sub-criteria based on this foundation. Each set of task-specific sub-criteria then guides the creation of well-detailed rubrics with specific questions tailored to each user prompt or problem statement, ensuring a nuanced approach to content evaluation. For example, in the task of poem writing, sub-criteria such as 'Sound and Rhythm', 'Structure and Form' etc reflect the unique aspects of poetic expression. Similarly, for blogs, sub-criteria like 'Clarity and Structure', 'Engagement' etc are designed to capture the analytical depth and reader involvement essential for effective blog writing. Each task-specific sub-criterion is operationalized through specific questions within the rubrics designed for individual problem statements. For instance, a problem statement about writing a reflective poem on a farmer's struggle includes a question under the sub-criterion 'Adherence

to Theme': 'Did the poem effectively address the theme of a farmer's dedication during tough times?' This question is assessed using a rating scale that evaluates the depth of perseverance and hope portrayed. Similarly, there are specific questions designed for each problem statement corresponding to the task-specific sub-criteria, ensuring that every aspect of the content is meticulously evaluated. This structured approach is done to ensure a comprehensive and consistent evaluation of content, enabling detailed comparison across models and tasks.

3.3. Large Language Models

We consider a diverse array of LLMs to evaluate their performance across various content generation tasks, involving both proprietary and open-source models. Our study evaluates 11 LLMs, including proprietary models such as OpenAI's GPT-3.5, GPT-4 [18] , Anthropic's Claude 2[19] , and Google's Gemini-Pro [20]. It also includes open-source models like meta-llama/Llama-3-70b-chat-hf [21], google/gemma-7b[22], mistralai/Mixtral-8x7B-Instruct-v0.1[23], Qwen/Qwen1.5-72B-Chat [24], Qwen/Qwen1.5-7B-Chat, meta-llama/Llama-2-7b-chat-hf[25], and lmsys/vicuna-13b-v1.5.By analyzing these models, we aim to determine if proprietary models are superior or if open-source models of various sizes are also competent enough to perform well on different writing tasks.

3.4. Prompting Techniques

We investigate the impact of different prompting techniques on the output quality of Language Models (LLMs) for various writing tasks, employing both "User prompts" and "System prompts." "User prompts" are direct problem statements specifying the content the model needs to generate, such as creating a poem or writing a blog. In contrast, "System prompts" are guiding instructions or contexts that direct the model's approach, like adopting a specific persona or style. We explore three primary prompting techniques: Zero-Shot Prompting, where only the user prompt is provided, allowing us to assess the LLM's intrinsic ability to interpret and respond without additional guidance and evaluate its baseline capability to produce coherent and contextually appropriate content. Contextual Role-Induced Prompting, which enhances the zero-shot method by adding a system message that defines a specific role or context for the LLM, such as a "creative poet," to guide the model's creative process, ensuring adherence to thematic and stylistic demands, thus enhancing the depth and quality of the generated content, and Chain of Thought (CoT) Prompting, a structured approach that makes the LLM engage in a two-phase process of planning and execution, where a system message first encourages the model to methodically plan its response before generating the content, fostering a disciplined and systematic content generation process and promoting clarity, inspired by the "Tree of Thought" [26] paper, to evaluate the model's ability to plan and execute tasks effectively.

3.5. Multi-Agent and Retrieval-Augmented Generation (RAG) Approaches

We explore the multi-agent framework and Retrieval-Augmented Generation (RAG) [27] to understand the enhancement in content generation. In campaign or ad video scriptwriting, specialized agents focus on specific elements such as plot writing, character development, taglines, scene description, and dialogue. We analyze if the collaboration of specialists improves

the content. For news generation, we use a RAG approach combined with a multi-agent framework to produce accurate and up-to-date articles. A News Collector agent gathers the latest information, while an Article Writer agent synthesizes it into a coherent piece. This method is applied specifically to closed-source LLMs, as open-source models do not support function calling, which is essential for integrating real-time data. This setup assesses whether models accurately use the obtained data for facts or introduce inaccuracies.

3.6. Evaluation Methodology

We plan to use LLMs to evaluate the content because we want to understand their effectiveness in assessing creative writing tasks. Two studies used GPT-4 for evaluating generated content and stories but did not justify why they chose GPT-4 over other models like Claude 3, Gemini 1.5 Pro, and Llama 3 70B. Additionally, no benchmarks exist for evaluating LLMs' performance in content writing tasks such as poems, blogs, and stories.

3.6.1. Selection of evaluator LLM:

We referred to the recent work in [7] which evaluated the quality of stories generated by 12 LLMs and 5 human writers. Ten honors and postgraduate Creative Writing students rated these stories using a detailed rubric. We used this corpus and expert ratings as a benchmark to select the best content evaluator LLM. We considered 4 LLMs—GPT-4, Claude 3, Gemini 1.5 Pro, and Llama 3-70B to rate all 65 stories from the corpus using the same rubric used by the human experts, ensuring unbiased evaluation by not disclosing the generator LLM of each story. Their study demonstrated a variance of up to 30% in scores assigned by different human evaluators, while consistently distinguishing between low-quality and high-quality content. We managed a similar setup, considering two temperatures for each evaluator LLM: t=0 and t=1. For each evaluator LLM, we calculated the average rating and standard deviation for the 13 candidate story writers across the 5 sets and compared these metrics to human expert ratings using the Euclidean distance formula given below.

$$\text{distance} = \sqrt{(x_{\text{LLM}} - x_{\text{human}})^2 + (y_{\text{LLM}} - y_{\text{human}})^2}$$

where x_{LLM} and y_{LLM} represent the average score and standard deviation of the LLM's ratings, and x_{human} and y_{human} represent the average score and standard deviation of the human expert ratings. This determined each evaluator LLM's proximity to the human benchmark.

The results, visualized in a graph 2, show average story ratings and standard deviations for each model. The Euclidean distance calculations, presented in the table 1, indicate that Llama 3-70B and Gemini 1.5 Pro have the closest alignment with human ratings. These results highlight that Gemini 1.5 Pro and Llama 3-70B are the most similar to human evaluators, making them strong candidates for content evaluation tasks.

We utilized Gemini 1.5 Pro, Llama 3-70B, and GPT-4 for evaluating poems, blogs, and ad copies, ensuring a comprehensive assessment of how these LLMs understand and rate high-quality and low-quality content. Each piece was evaluated at temperature settings of 0 and 1 to capture scoring variability. For news article evaluation, we exclusively used GPT-4 due to its real-time

Table 1
D is the average Euclidean distance between the ratings given by each evaluator LLM and the human ratings, indicating the proximity of each LLM's ratings to the human benchmark. The table includes distances (D) for temperature 0 and 1 settings, as well as their average(D - overall).

Model	D for t=0	D for t=1	D - overall
Claude 3	23.69	19.47	22.24
GPT-4	17.29	19.68	18.84
Llama 3	12.44	13.93	13.50
Gemini	14.98	11.84	13.53

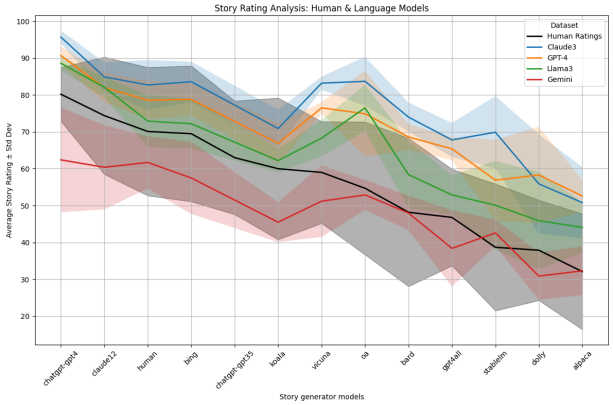


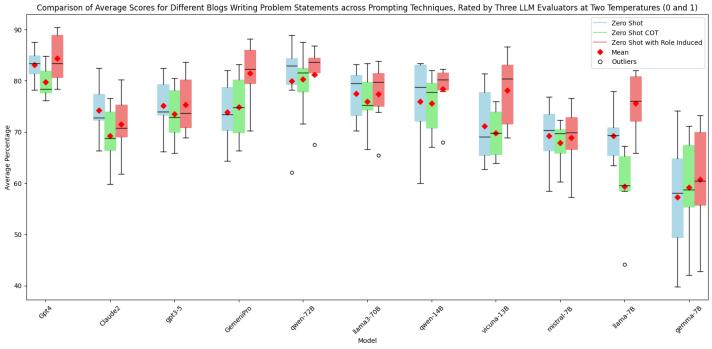
Figure 2: Comparison of story ratings by humans and 4 LLM's. The plot shows the average story ratings with standard deviations, illustrating the alignment of each LLM's ratings with human ratings.

internet access and superior fact-checking capabilities, ensuring accuracy and relevance. For short movie script generation, Gemini 1.5 Pro and Llama 3-70B served as evaluators.

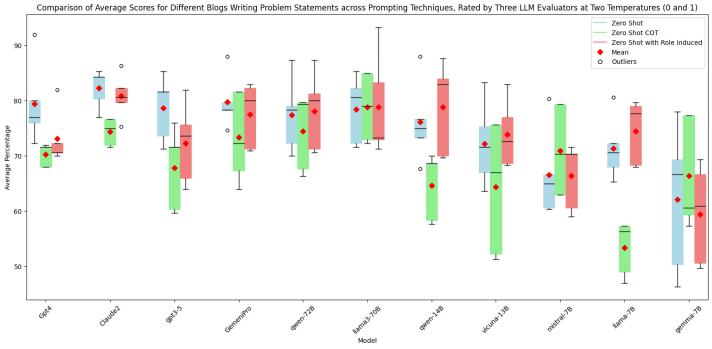
Due to a lack of expert evaluators, formal human evaluations were not conducted. However, to understand human preferences for AI-generated versus human-generated content, we conducted a survey using Google Forms with 50 participants. This survey aimed to gauge human preferences rather than serve as a formal evaluation. Without disclosing whether the content was human or AI-generated, all participants were presented with the same form containing two poems, two blogs, and two news articles. Each type included one human-written and one LLM-generated piece addressing the same problem statement. Participants were asked to select their preferred creation for each set, allowing for a direct comparison of preferences across different content types.

4. Experiment Setup

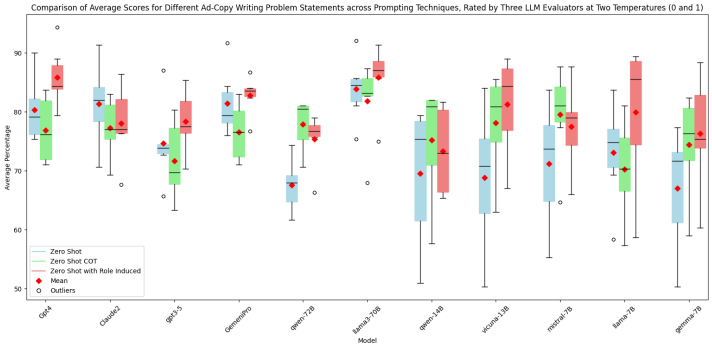
For our experiments with open-source LLMs, we use the Together AI service, while interactions with GPT-4 and GPT-3.5 are conducted via the OpenAI Playground. Claude is accessed through the Anthropic Playground, and Gemini-Pro and Gemini-1.5-Pro are utilized through Google AI Studio. For implementing agent frameworks, we develop an agent team using CrewAI [28], utilizing the API services provided by the aforementioned platforms. Additionally, for news



(a) Poems



(b) Blogs



(c) Ad-Copies

Figure 3: Box plot displaying performance scores of LLMs across three writing tasks (blog, poem, and ad copy writing) and three prompting strategies (Zero Shot, Zero Shot with Role Induced, Zero Shot COT). Evaluations were conducted using three LLMs (Llama-3, GPT-4, and Gemini 1.5-Pro) at temperature settings 0 and 1. For each problem statement, scores from different rubric items are totaled per evaluator LLM, then these totals are converted to percentages and combined across all three LLMs for both temperature settings. Boxes in the plot represent the median and quartiles (Q1-Q3), while whiskers extend to 1.5 times the interquartile range (IQR). Outliers and means are indicated as black hollow circles and red diamonds, respectively. This plot highlights the comparative effectiveness of prompting strategies and performance variability between models and within models across tasks.

searches, we utilize the DuckDuckGo[29] service to fetch articles from the internet. The CrewAI team was configured to operate sequentially for news search tasks and hierarchically for movie script generation.

We consider three to four problem statements for every writing task, employing three types of prompting methods, and execute each problem statement five times per prompting type. This approach ensure consistency and diversity in understanding instructions. Consequently, we generated 660 poems, 495 blogs, and 495 ads across 11 LLMs. However, due to resource limitations and infrastructure costs, only one-fifth of these outputs (132 poems, 99 blogs, and 99 ads) were evaluated. Each piece of content was assessed by GPT-4, Gemini 1.5 Pro, and Llama 3-70B, with the evaluators set at temperature values of 0 and 1, resulting in six evaluations per content piece. For movie script generation, the six LLMs used were GPT-4, GPT-3.5, Claude 2, Gemini-Pro, Qwen 1.5-72B, and Llama 3-70B, producing 18 scripts evaluated by GPT-4, Gemini 1.5 Pro and Llama 3-70B. The multi-agent setup for movie scripts and news articles utilized the CrewAI framework, where the agents were ReAct-based agents . In the news generation task, four LLMs—GPT-4, GPT-3.5, Claude 2, and Gemini-Pro—produced 12 articles, which were evaluated exclusively by GPT-4 due to its web searching capabilities. We calculate BERT scores [30] within the five iterations of running each prompt in each setting to assess the diversity of the outputs, determining whether the LLMs were generating varied content or if there were consistent core ideas across the iterations.

5. Results and Discussion

Our analysis reveals some good insights into the performance of various LLMs across multiple creative writing tasks.

5.1. LLMs as content writers

Across various writing tasks, LLMs show significant variability in content quality, topic adherence, clarity, and emotional impact. GPT-4 and Qwen 1.5-72B excel in thematic consistency, emotional depth, sound, and rhythm in poems. Claude 2 produces well-structured and clear content but sometimes refuses to generate advertisements or handle sensitive topics. Llama 3-70B maintains clarity, persuasion, and flow in blogs and ads, and also performs well in poems. Gemma-7B struggles significantly across all tasks, failing in stanza structures, language appropriateness, and clarity, resulting in free-form text and incoherent outputs. Mixtral-8x7B, Vicuna-13b, and Llama-2-7b generally adhere to problem statements but sometimes struggle with maintaining clarity in intricate themes or longer formats. Qwen 1.5-7B-Chat performs moderately well but occasionally produces content in Chinese, affecting clarity. Gemini Pro excels in blogs and ads, maintaining structural integrity and clarity, but falls short in sound, rhythm, and emotional impact in poems. GPT-3.5 shows moderate adherence to problem statements and quality in content generation for poems, ads, and blogs but struggles with maintaining consistent clarity and coherence. For blog writing, most LLMs fulfilled the word limit requirements for shorter texts (less than 500 words), but all struggled with higher word limits like 2000 words. Additionally, almost all LLMs failed to count words accurately, indicating a limitation in word

count accuracy. However, Claude 2, Llama 3-70B, and Qwen 1.5-72B generated longer contexts more effectively than others.

5.2. Effect of Prompting Strategies

In poem writing, the Zero Shot Role Induced prompting consistently outperforms other strategies, significantly enhancing thematic consistency, emotional depth, and structural elements such as sound and rhythm. This strategy benefits models like GPT-4, Llama 3-70B, and Qwen 1.5-72B the most, as illustrated in 3a , where these models show marked improvements in performance scores. For ad copy generation, Zero Shot Role Induced prompting is the leading strategy, improving clarity, engagement, and visual appeal, with GPT-4, Llama 3-70B, and Claude 2 showing the most significant gains, as depicted in 3c. In blog writing, both Zero Shot and Zero Shot Role Induced prompting perform well, with GPT-4 and Llama 3-70B excelling in clarity and structural coherence. However, Chain of Thought (CoT) prompting in blog writing performs poorly as it consumes many tokens for planning, resulting in shorter and less comprehensive outputs, which can be observed in the comparative effectiveness of prompting strategies shown in 3b. Smaller models like Gemma-7B and Llama-2-7b often fail to plan content properly, leading to inconsistent outputs, while models like Vicuna-13b, Mixtral-8x7B, and Qwen 1.5-7B-Chat sometimes struggle with executing their plans effectively.

5.3. Factual Consistency in News

For news article generation, our analysis by manual observation indicates that GPT-3.5 occasionally introduced extraneous facts when the internet search by the news searcher agent was inadequate. In contrast, Claude 2 exhibited a different behavior: if the news searcher agent did not gather sufficient information, the writer agent would refuse to perform the writing task due to the lack of necessary data. On the other hand, both Gemini Pro and GPT-4 demonstrated overall better performance, integrating accurate and relevant information into the news articles.

5.4. Multi-Agent Collaboration

Regarding movie script generation, GPT-4, Llama 3-70B, and Qwen 1.5-72B slightly outperformed other models in both single and multi-agent settings, as shown in Figure 4. However, Claude2 agents, despite seeking extensive information about plots, characters, and scenes, did not show a significant impact on output quality. GPT-3.5, on the other hand, showed a noticeable decline in performance when operating as a single agent compared to its multi-agent setup, indicating a potential reliance on collaborative workflows to optimize output.

A notable shortfall across all models was their lack of creativity and out-of-the-box thinking, often reflecting stereotypical mindsets. For example, in generating SnoreAway ad scripts, all models predominantly depicted men as the snorers, overlooking women. The generated plots, scenes, and dialogues were generally generic and lacked expert-level imagination. Interestingly, despite the general superiority of multi-agent setups as shown in the figure, experiments revealed that in some cases, such as with GPT-4 and Llama 3-70B, the single agent's performance was relatively competitive, showing only a slight drop of about 5-15% in scores compared to their multi-agent counterparts. This suggests that under certain conditions, single zero-shot agents

can nearly match the performance of multi-agent setups, offering a resource-efficient alternative with lower token consumption. This observed efficiency warrants further investigation into whether optimizing single agents for multi-task roles could reduce resource consumption without compromising the quality of the output.

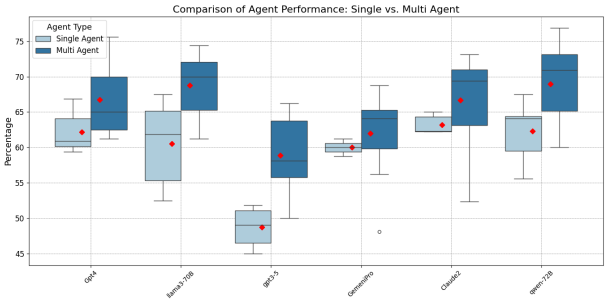


Figure 4: Box plot summarizing AI model performance under single and multi-agent settings. The plot displays median, quartiles (Q1-Q3), and 1.5 IQR whiskers, with outliers as black hollow circles. Red diamonds mark the means. Evaluations involved two LLMs (Llama-3 and Gemini 1.5-Pro) at temperature settings 0 and 1.

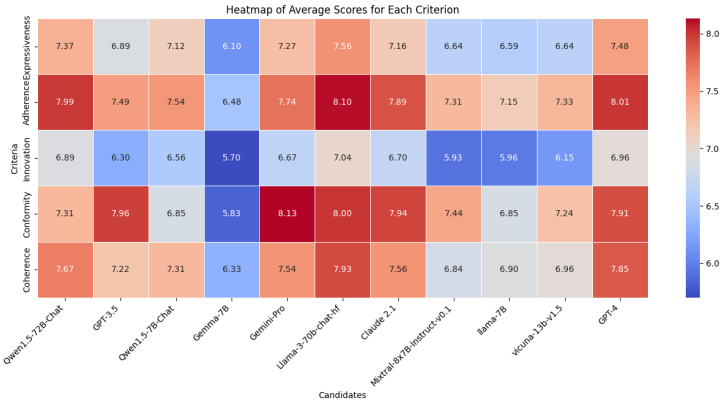


Figure 5: Heatmap illustrating summarized performance of LLMs by aggregating scores from a bottom-up approach within our evaluation hierarchy (as described in 3.2). Scores start from specific questions addressing task-specific sub-criteria for all tasks and models across two temperatures. These are then aggregated to represent each task-specific sub-criterion, and finally compiled into the high-level criteria, illustrating the comprehensive performance across all tasks.

The heatmap analysis 5 highlights that GPT-4 and Llama3-70B are the overall best-performing LLMs, while Gemma-7B consistently underperforms across tasks. The evaluators, LLMs themselves, rated high-quality content positively and low-quality content negatively, demonstrating their potential for reliable content evaluations. Notably, using multiple evaluator LLMs enhances the reliability of the results. As depicted in the figure 5, innovation is the lowest-rated criterion, indicating that the generated content lacked originality and out-of-the-box thinking, tending to be basic and easily reproducible. For measuring diversity, we have calculated BERT scores across the five generations made for each problem statement to understand the dissimilarity of outputs

for the same prompt in the same setting. Higher dissimilarity values, as seen with Llama 3-70B (19.23%) and Mistral-7B (17.94%), indicate that these LLMs are capable of generating diverse and novel content pieces for the same user prompt. Other models show dissimilarity in the range of approximately 11% to 17.5%. The overall performance of open-source models ranged from average to good, suggesting that they are viable options for creative tasks. Additionally, the LLMs showed a consistent ability to adhere to structural and stylistic guidelines, indicating their proficiency in producing coherent and well-organized content. This suggests that while they may not yet excel in innovative thinking, they are competent in following detailed instructions and maintaining clarity and structure in their outputs. In our survey taken for human preference, as mentioned in 3.6, for poems, respondents tended to select the version they found easier to understand. For the first poem's problem statement, 52.9% favored the AI-generated content, appreciating its clarity, while 53.8% preferred the human-written version of the second poem's problem statement for its simplicity. For blogs, preference leaned towards LLM-generated content, with 62.7% favoring AI for the first blog and 58.8% for the second. Similarly, 54.9% preferred AI-generated news articles for both problem statements, valuing their clarity and engagement. Even though the differences are not significant, AI-generated content is equally preferred and often chosen alongside human-created content. These findings suggest that content clarity significantly drives preferences, with respondents favoring clearer presentations from both humans and LLMs.

6. Conclusion

Our study effectively demonstrates the capabilities and limitations of various LLMs across multiple creative writing tasks. Our comprehensive evaluation highlights the strengths of models like GPT-4 and Llama 3-70B, while also identifying areas for improvement, particularly in fostering creativity and originality. Additionally, the LLMs showed a consistent ability to adhere to structural and stylistic guidelines, indicating their proficiency in producing coherent and well-organized content. This suggests that while they may not yet excel in innovative thinking, they are competent in following detailed instructions and maintaining clarity and structure in their outputs. Interestingly, our results indicate that proprietary models are not always significantly superior, with open-source models like Qwen and Llama-3 outperforming GPT-3.5 and Gemini in several tasks, demonstrating that mid-ranged open-source models can perform competitively. The use of multiple LLMs as evaluators has proven reliable for content assessment, emphasizing the potential of automated evaluations. However, the study is limited by the absence of human-based evaluations and reliance on human preferences. Future research should include more diverse prompting techniques, explore alternative content evaluation strategies, and assess advanced Retrieval-Augmented Generation (RAG) systems beyond the classic RAG methodology employed here. Additionally, the potential of single-agent setups versus multi-agent frameworks warrants further investigation to optimize content quality across multiple subtasks.

7. Acknowledgments

We thank Mr. Hari Narayan, a researcher at TCS, for his valuable ideas, insights and reference implementation of the LLM-as-Judge-based content evaluation experimental setup.

References

[1] M. Donovan, M. Donovan, Types of creative writing | Writing forward, 2021. URL: https://www.writingforward.com/creative-writing/types-of-creative-writing#google_vignette.

[2] R. Lowe, Different writing styles: Exploration of 9 powerful artistic forms, 2024. URL: https://thewritingking.com/different-writing-styles/.

[3] K. Cummins, Text Types and Different Styles of Writing: The Complete guide, 2024. URL: https://literacyideas.com/different-text-types/#poetry.

[4] D. Zelikman, 10 Types of Creative Writing (with Examples You'll Love), 2021. URL: https://blog.reedsy.com/guide/creative-writing/types-and-examples/.

[5] T. Chakrabarty, V. Padmakumar, F. Brahman, S. Muresan, Creativity support in the age of large language models: An empirical study involving emerging writers, ArXiv abs/2309.12570 (2023). URL: https://api.semanticscholar.org/CorpusID:262217523.

[6] Z. Xie, T. Cohn, J. H. Lau, The next chapter: A study of large language models in storytelling, in: C. M. Keet, H.-Y. Lee, S. Zarrieß (Eds.), Proceedings of the 16th International Natural Language Generation Conference, Association for Computational Linguistics, Prague, Czechia, 2023, pp. 323–351. URL: https://aclanthology.org/2023.inlg-main.23. doi:10.18653/v1/2023.inlg-main.23.

[7] C. Gómez-Rodríguez, P. Williams, A confederacy of models: a comprehensive evaluation of LLMs on creative writing, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 14504–14528. URL: https://aclanthology.org/2023.findings-emnlp.966. doi:10.18653/v1/2023.findings-emnlp.966.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.

[9] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, 2023. arXiv:2210.03629.

[10] C. Lamb, D. G. Brown, C. L. A. Clarke, Human competence in creativity evaluation, in: International Conference on Innovative Computing and Cloud Computing, 2015. URL: https://api.semanticscholar.org/CorpusID:14806090.

[11] G. Franceschelli, M. Musolesi, Creativity and machine learning: A survey, ACM Computing Surveys (2024). URL: http://dx.doi.org/10.1145/3664595. doi:10.1145/3664595.

[12] F. Pereira, M. Mendes, P. Gervás, A. Cardoso, Experiments with assessment of creative systems: An application of ritchie's criteria, 2005.

[13] A. Jordanous, A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative, Cognitive Computation 4 (2012). doi:10.1007/s12559-012-9156-1.

[14] S. Colton, J. Charnley, A. Pease, Computational creativity theory: The face and idea

descriptive models, in: Proceedings of the 2nd International Conference on Computational Creativity, ICCC 2011, Proceedings of the 2nd International Conference on Computational Creativity, ICCC 2011, 2011, pp. 90–95. URL: https://computationalcreativity.net/iccc2011/proceedings/index.html, international Conference on Computational Creativity 2011, ICCC 2011 ; Conference date: 27-04-2011 Through 29-04-2011.

[15] M. A. Boden, Understanding creativity, J. Creat. Behav. 26 (1992) 213–217.

[16] Y. Zhao, R. Zhang, W. Li, D. Huang, J. Guo, S. Peng, Y. Hao, Y. Wen, X. Hu, Z. Du, Q. Guo, L. Li, Y. Chen, Assessing and understanding creativity in large language models, ArXiv abs/2401.12491 (2024). URL: https://api.semanticscholar.org/CorpusID:267094860.

[17] T. Chakrabarty, P. Laban, D. Agarwal, S. Muresan, C.-S. Wu, Art or artifice? large language models and the false promise of creativity, 2024. arXiv:2309.14556.

[18] O. Team, Gpt-4 technical report, 2024. arXiv:2303.08774.

[19] Introducing the next generation of Claude Anthropic, ???? URL: https://www.anthropic.com/news/claude-3-family.

[20] G. Team, Gemini: A family of highly capable multimodal models, 2024. arXiv:2312.11805.

[21] W. Huang, X. Ma, H. Qin, X. Zheng, C. Lv, H. Chen, J. Luo, X. Qi, X. Liu, M. Magno, How good are low-bit quantized llama3 models? an empirical study, 2024. arXiv:2404.14047.

[22] G. Team, Gemma: Open models based on gemini research and technology, 2024. arXiv:2403.08295.

[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[24] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, T. Zhu, Qwen technical report, 2023. arXiv:2309.16609.

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[26] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, 2023. arXiv:2305.10601.

[27] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. arXiv:2312.10997.

[28] Joaomdmoura, GitHub - joaomdmoura/crewAI: Framework for orchestrating role-playing, autonomous AI agents. By fostering collaborative intelligence, CrewAI empowers agents to work together seamlessly, tackling complex tasks., ???? URL: https://github.com/joaomdmoura/CrewAI.

[29] duckduckgo-search — pypi.org, https://pypi.org/project/duckduckgo-search/, ???? [Accessed 29-05-2024].

[30] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. arXiv:1904.09675.