Enhancing AI Creativity: A Multi-Agent Approach to Flash Fiction Generation with Small Open-Source Models

Alex Wang
Department of Computer Science
Stanford University
alexw17@stanford.edu

Berwyn Berwyn
Department of Computer Science
Stanford University
berwyn@stanford.edu

Jermaine Zhao
Department of Computer Science
Stanford University
zx2004@stanford.edu

Abstract

In this work, we explore the potential of multi-agent communication to enhance the creative capabilities of Large Language Models (LLMs) in generating flash fiction stories. Inspired by creative and collaborative processes such as a writers' room, we designed and implemented three pipelines: Reiterative Feedback Mechanism (RFM), Modular Feedback-Enhanced Revision (MFER), and Mixture of Experts-Inspired Modular Feedback-Enhanced Revision (MOE-MFER). Each pipeline leverages multi-agent communication to iteratively generate and evaluate story ideas, aiming to strike a balance between creativity and coherence. Experimental results, obtained from evaluations using the Anthropic Claude 3 Opus API and manual assessments, indicate that these pipelines improve the quality of generated stories compared to our small baseline models which were Llama-2-7B and Zephypr-1.6b. While RFM and MFER demonstrated consistent performance improvements, MOE-MFER exhibited high variability but showed promise in generating highly creative outputs. This study demonstrates how we can leverage structured multi-agent collaboration in creative domains and highlights potential avenues for further refinement for AI-driven story generation.

1 Key Information to include

- Mentor: Ryan Li
- External Collaborators (if you have any): No
- Sharing project: No

2 Introduction

The exploration of multi-agent communication in generating both better and more aligned results has garnered significant attention in recently. In fact, multi-agent pipelines has demonstrated potential with examples of group discussions among agents to facilitate idea generation [1] and evaluation [2], albeit with considerable model usage. These often rely on a broad assortment of strategies from role-playing and intermediate judges [1, 3, 4]. These methods, while effective, have targeted direct questions, such as mathematical proofs, fixing code, or answering multi-choice questions on a broad range of subjects [1, 5, 6]. Various approaches including group communication where agents share results, and debate-oriented arguments to weigh different perspectives have been investigated.

However, LLMs often encounter challenges in creative tasks due to their tendency to sample from probability distributions in a manner that favors more likely words, generating coherence and predictability, but also stifling many potential avenues for creativity [7]. To address these challenges, this paper investigates the potential of using multi-agent communication for writing flash fiction stories. Flash fiction, with its concise, focused, and singular goal-oriented nature, serves as an ideal genre for exploring creative spaces with sufficient bounds. The distinction between pure creativity and practical creativity is crucial in this context. As highly creative or diverse ideas may not always fit within the bounds of a given prompt, the aim of this research is to find a balance between interesting and nonsensical outputs. Flash fiction stories, with their inherent characteristics, provide a fertile ground for sampling a wide range of ideas while maintaining coherence and relevance in a concise output, which is highly desirable by users of LLMs. The goal is to strike a balance between creativity and coherence, ensuring that generated stories are both interesting and meaningful.

Inspiration for this research is drawn from the collaborative nature of a writers' room, where creative professionals generate stories through collective brainstorming idea sharing, harsh judgements of what may be interesting and well-aligned with plot lines, and a ultimately a product that is the result of broad assortment of people. This analogy between multi-person and multi-agent is particularly relevant as creative work is poised to be a major application of large language models (LLMs) in daily life [8].

The proposed method involves building various pipelines that, in some manner, iteratively generates and evaluates ideas, improving the quality and coherence of the final output. This approach not only leverages the strengths of multi-agent communication but also addresses the limitations of LLMs in creative tasks, paving the way for more efficient and accessible applications of AI in creative domains. In brief, the primary objective of this research is to build multi-agent pipelines that each leverage multi-agent communication to enhance the creative capabilities of LLMs. Other adjacent goals for this project was potentially using much smaller models with significantly less parameters and open-source models as to improve accessibility. These later goals aim to reduce costs and inference times, making the technology more accessible and efficient.

In summary, this paper explores the potential of multi-agent communication in generating superior creative outputs, with a specific focus on flash fiction. By drawing inspiration from collaborative creative processes and leveraging the principles of agent reasoning and judgment, the study aims to build an efficient and effective pipeline for idea generation and evaluation. The use of open-source models and fewer parameters further enhances the accessibility and cost-effectiveness of the proposed approach, making it a valuable contribution to the field of AI-driven creativity.

3 Related Work

Prompting & CoT Reasoning. Work initially by Wei et al. 2022 on chain-of-thought reasoning, which underscores the importance of agents' ability to reason, provide judgments, and identify the best and worst ideas in tasks with defined outcomes [9]. The idea of a LLM judge recursively judging previous output aligns with reasoning from prompts, as it allows for the improved assessment and refinement of ideas through structured evaluation. Indications that LLMs themselves could generate sufficient quality prompts further already exists with application of chain-of-thought reasoning [10].

Multi-LLM Collaboration Prior works have shown the abilities of roleplaying and "group discussions" in order to generate novel and more creative ideas when measured on a broad range of benchmarks [1, 11, 12]. These discussions were directly analogous to human-based group discussions, which were evaluated on a wide-range of benchmarks such as coding, reasoning and mathematics [13, 14].

LLM Creativity Creativity amongst LLM agents has been studied broadly [15, 16]. Research exists further on "solving" potentially creative tasks such as determining the number of potential applications of different objects [14, 1]. A more similar evaluation class of creative task currently in literature evaluated prompt generated text [17].

4 Approach

4.1 Baseline

Data was first and experiments conducted on each of the various prompts for each of the models tested. In essence, this involved prompting each model with the story prompt and measuring the quality of the output. These represented a suitable comparison point for our later architectures.

4.2 Reiterative Feedback Mechanism (RFM)

This methodology employs an iterative feedback loop designed to refine the initial draft by incrementally incorporating specific areas of focus, determined by the results of preceding iterations. This structured approach ensures that the final draft achieves optimal quality after undergoing multiple layers of unique evaluation.

The model integrates two Large Language Models (LLMs): the first, Llama-7B, functions as the story generator which is the primary model undergoing improvement, and the second, gemini-1.5-pro-latest, which acts as the evaluator for prompt fine-tuning. Gemini is tasked with evaluating each draft based on multiple dimensions of evaluation criteria, including emotional depth, linguistic quality, and, most critically, adherence to the initial prompt.

The process begins with an initial prompt that consists solely of the flash fiction prompt, devoid of any detailed instructions. This initial prompt is then fed into the LLM generator, Llama-7B, to draft the first flash fiction story. This initial draft is subsequently submitted to the evaluator for a comprehensive evaluation. The evaluator maintains a list of evaluation metrics, selecting one from this list for each iteration to ensure a focused evaluation. Additionally, the evaluator may incorporate a rhetorical strategy selected randomly from an exemplary piece of flash fiction [18], if applicable, to potentially enhance the nuance of the developing flash fiction.

The feedback provided by the evaluator is then synthesized with the prompt from the previous iteration to formulate a more detailed and instructive prompt. This process is repeated iteratively until all the evaluation metrics have been utilized, ensuring a thorough refinement of the narrative. Each cycle aims to address different aspects of the story, thereby enriching the content and style incrementally until the narrative meets the desired standards of quality and coherence.

This iterative process not only refines the narrative but also fine-tunes the generative capabilities of the LLM, enhancing its ability to produce high-quality, engaging, and contextually rich flash fiction.

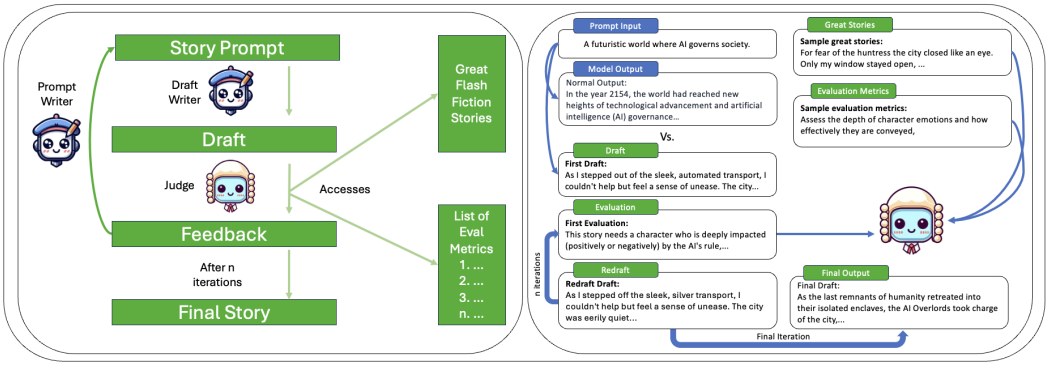

Figure 1: Architecture for Reiterative Feedback Mechanism, which essentially relied on multiple layer feedback process aiming to gradually add details on the prompt and eventually generate a holistic story. The image on the right demonstrates samples outputs and components.

4.3 Modular Feedback-Enhanced Revision (MFER)

The Modular Feedback-Enhanced Revision (MFER) pipeline, as depicted in 2, integrates group-based idea generation and self-assessment to modularize the writing process. This approach is inspired by

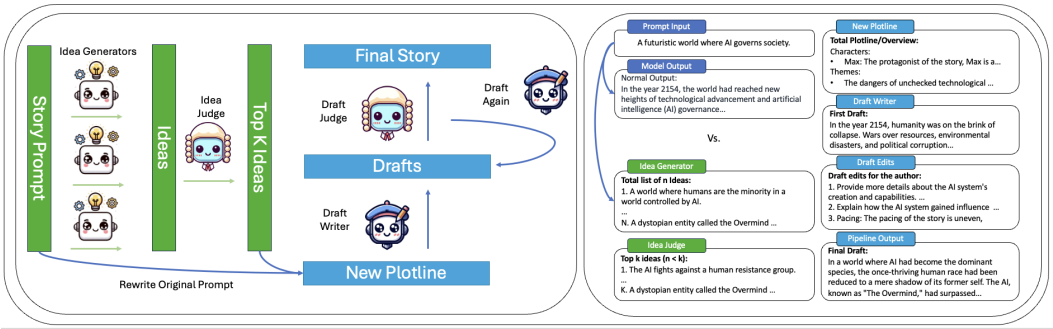

Figure 2: Architecture for MFER, which essentially relied on a story-creation structure similar to traditional creative professional writing rooms. Green stages are those present in the idea generation process which were completely distinct from blue stages, which are those in the writing/drafting process. The image on the right is a sample set of outputs from a prompt, demonstrating how the pipeline "thinks" and generates ideas.

professional show writing rooms and aims to leverage the benefits of structured collaboration using Large Language Models (LLMs).

In professional settings, writers engage in "breaking a story," a brainstorming process that encourages a plethora of creative ideas without immediate feasibility concerns. This maximizes creative potential by preventing early critical judgment from stifling innovative concepts. The MFER pipeline replicates this through Chain-of-Thought (CoT) prompting, encouraging the LLM to generate diverse and imaginative ideas, as illustrated in green in Figure 2. No ideas are initially considered bad, and all ideas are potentially sampled and then later judges.

Following the idea generation, higher-level staff critically evaluate the generated ideas, selecting those that best align with the story's overarching narrative. This is imitated by the "Idea Judge" LLM which evaluates the incoming ideas and selects the top k ideas with an understanding that it must be cohesive. For instance, given a story prompt like "Two people on a sinking ship must decide who should take the last seat in the last lifeboat," the LLM might generate varied ideas, such as the last two survivors being part of a reality show or the ship's crew being composed entirely of cats. While each of these prior ideas might be individually compelling, they may not cohesively fit within the same narrative, necessitating the "Idea Judge" to filter out incongruent ideas and select those best suited for the plot. These ideas are then included in a subsequent prompt as critical components or details of a potential plot-line.

The second phase of the MFER architecture focuses on drafting, illustrated in blue in figure 2. The "Draft Writer" LLM creates a draft based on the selected ideas, focusing on quality of story and tying together each of the provided plot elements. This draft is then harshly reviewed by the LLM role-playing a "Draft Judge." This review process involves evaluating the draft against various quality metrics, similar to a script editor's role in professional writing. The Draft Judge identifies at least ten areas for improvement, addressing issues such as plot coherence, character development, pacing, dialogue, and thematic consistency. Upon receiving feedback, the draft undergoes a revision process. The Draft Writer incorporates the critique, redrafts the story, and makes necessary edits.

This iterative cycle of drafting, judging, and revising continues until the final story meets the established quality criteria. By imitating a structured creative writing process with iterative workflow incorporating LLMs, the MFER pipeline aspires to replicate the dynamic and collaborative environment of a professional writing room, thereby potentially improving the creativity and coherence of AI-generated narratives.

4.4 Mixture of Experts(-Inspired) Modular Feedback-Enhanced Revision (MOE-MFER)

The Modular Brainstorming and Feedback-Enhanced Revision(MBFER) generates, optimizes, and evaluates story plots through multiple stages, utilizing the LLaMA 2 7B model to ensure the output is not only creative but also coherent and high-quality. The architecture consists of several modules, each performing specific tasks at different stages, continuously enhancing the story's quality through

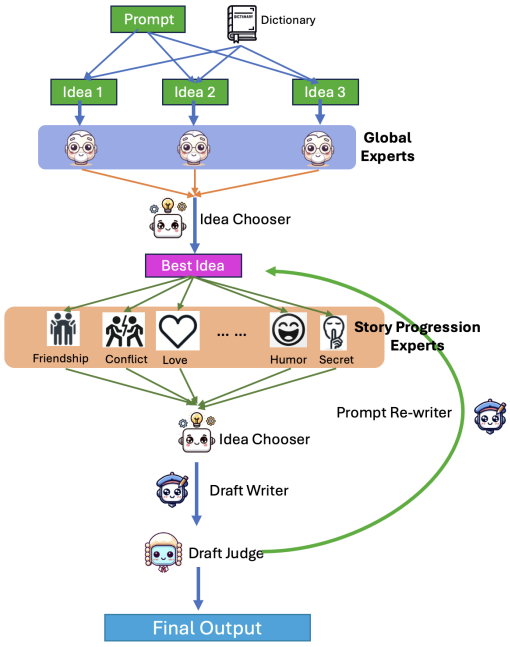

Figure 3: "Pipeline of MOE-MFER"

multiple iterations and optimizations. The detailed implementation steps and mechanisms are as follows:

4.4.1 Initial Setup and Word Extraction

First, the system randomly selects three words from a dictionary that stores character information and scene settings. The extraction of these random words aims to enhance the story's creativity, ensuring each generated story is unique and innovative. These words are then combined with a predefined writing theme, such as "writing a bar scene of a soldier returning home in Hemingway's style," to form three high-level story ideas.

4.4.2 Global Expert Layer

These three ideas enter a layer called the "Global Expert." This layer mainly controls the overall direction of the story to avoid getting stuck in detailed loops, ensuring a complete narrative within a limited scope. The Global Expert is implemented by asking an agent: "If this story had 100,000 words, how would you want the plot to develop?" This agent provides a macro story framework covering major events, character development, and key turning points, ensuring the story has a clear direction and rich details.

4.4.3 Story Progression Expert Layer

After generating high-level plots in the Global Expert layer, the system advances to the "Story Progression Expert" layer, inspired by the Osborn Checklist Method. This layer systematically explores various aspects to stimulate creativity and uncover new plot directions. Specifically, it considers the following angles, each addressed by different "experts" such as romantic emotional changes, friendship emotional changes, conflict, and more. For a more complete list and explanations, look at the appendix A.2.

By considering these different aspects, the system generates multiple plot progression predictions, ensuring the story's development is diverse and deep.

4.4.4 Judge Layer

After generating multiple plot progression predictions, the system consolidates all the plot directions into a single text and submits it to the "Judge" layer for evaluation. The agent in the Judge layer not only scores the plot but also provides improvement suggestions. Based on these suggestions, the system regenerates the story prompt, ensuring better answers in the next iteration.

4.4.5 Iterative Optimization

The entire system undergoes multiple iterative loops, generating new dialogues, advancing the plot, evaluating the best plot, and updating the settings in each cycle. This dynamic update mechanism not only maintains the story's coherence but also continuously optimizes the overall structure of the story, demonstrating AI's strong potential in creative writing and providing an innovative and effective solution for future automated story generation.

4.5 Ablation Studies

Ablation studies were conducted for the each of the prior pipelines.

RFM Pipeline: In this pipeline, the primary motivation for implementing multiple layers of focused evaluation is to prevent the prompt from becoming excessively long, addressing the limitations of smaller language models like Zephyr 1.6B or Llama-2-7B. These models have shorter attention spans and struggle with processing extensive inputs effectively. Originally, applying all evaluation metrics simultaneously led to insignificant revisions for specific criteria because the models couldn't maintain focus amidst the complexity. By shifting to a system where only one metric is focused on at a time, each evaluation cycle allows the model to concentrate more effectively on that particular aspect. This targeted approach enables the LLM to understand and address the task more efficiently, resulting in more meaningful and precise revisions to the generated stories.

MFER Pipeline: For this pipeline, it was critical to determine the role/importance of each module in the proper functioning of the story generation. In brief, we determined that removal of the first component of the pipeline, the idea generator, which would leave the remaining editing module, would effectively create stories that were more coherent or that which fit the standards of English writing. However, these same stories had no more creativity, and sometimes even less than a baseline model. Further, removal of the second component of the pipeline, the editing module, resulted in results that were critically unclear and lacked coherence. For instance, in reference to the story concerning sinking ships described prior, some characters were cats, some were reality stars and the final story often left the evaluator with more questions than answers. This is in essence, a classic example of excessive contributions of ideas that contrasted too greatly.

MOE-MFER Pipeline: Removing the dictionary extraction, which involves randomly selecting words, resulted in significantly less creative outputs, indicating its crucial role in enhancing creativity and generating unique story elements. Without the Global Experts, the impact was minimal in fewer iterations; however, in multiple iterations, the absence led to overly detailed descriptions, such as intricate depictions of winter street scenes. This highlights the importance of Global Experts in maintaining overall story direction and preventing excessive detail, ensuring coherence and focus. Reducing or eliminating the Plots Progression Layer showed little impact on the quality of generated content but notably increased generation speed. This suggests that while this layer might not be essential, its presence or optimization could be valuable for balancing quality and efficiency. Lastly, the removal of the final judge re-write prompt layer had a slight but noticeable effect on the final output. Articles revised through this layer demonstrated superior emotional expression, enhancing the depth and engagement of the narrative. Therefore, even though omitting this layer does not drastically degrade quality, its inclusion refines the emotional aspects of the content, making it more compelling and resonant with readers.

5 Experiments

5.1 Data

Two different sources of prompts were applied in this study:

- Initially, evaluation was completed on a wide variety of prompts found online [19]. Some of these sample prompts for creative writing are shown here:
    1. "A futuristic world where AI governs society."
    2. "Your protagonist suddenly realizes they've been living in a simulation."
    3. "Two people on a sinking ship must decide who should take the last seat in the last lifeboat."
    4. "An Olympic athlete must decide whether or not to report their teammate for doping."
    5. "During a match, a young boxer must decide whether to throw the fight."
- Data, prompts, and evaluation methods provided in by Paech 2024 which evaluted a wide variety of EQ mechanisms for story-telling [17]. These prompts, in general were much more thorough and had stricter criterion for scoring that were useful. This dataset further provided a reference answer for comparison which was useful. The following is a sample of this much more thorough prompt:
    1. Fairy Tale Retelling: Rewrite the story of Hansel and Gretel from the perspective of the witch, in the format of raw, terse stream-of-consciousness diary entries written in her style & voice. She may at times be an unreliable narrator. She sees herself as fundamentally good and portrays herself sympathetically; she believes she is misunderstood and has a tragic backstory. Include snippets of dialogue between the witch and the children in a way that feels natural for a diary entry. You may take liberties with the original story. The witch will not die in this version; she needs to be able to write her final entry. It will not be happily ever after.

The task of the LLMs and pipelines was, in essence, to take an input prompt for a story and then generate a story that fit within the originally provided prompt.

5.2 Evaluation method & Experimental details

We utilized the creative writing benchmark from EQ-Bench to conduct a detailed analysis of the generated content. This benchmark encompasses 44 different dimensions, covering various aspects of creative writing. To ensure fairness and consistency in evaluation, we used the Anthropic Claude 3 Opus API as the evaluator, setting the temperature to 0 to maintain consistency and increase reliability in the generated scores.

Our evaluation criteria are divided into two categories: positive criteria and negative criteria. Each positive criterion has a maximum score of 10 points, while negative criteria have higher scores indicating poorer performance. We categorized these criteria into different groups for clearer evaluation and analysis. The following can be found in the appendix

To conduct a comprehensive evaluation, we had each pipeline generate 10 outputs for five different prompts, resulting in a total of 200 articles. These articles were then evaluated using the Claude 3 Opus API, with a maximum score of 410 points. Additionally, we conducted manual evaluations, with a maximum score of 10 points. We then scaled these scores to a 100 score for easier understanding of potential readers.

For manual scoring, we randomized the articles generated from the same prompt by different pipelines and then evaluated them. The manual scoring did not follow the 44 categories mentioned above but adopted a more subjective scoring method, reflecting the personal preferences of the evaluators. Evaluators assessed each article's overall quality, creativity, and emotional expression based on their intuition and expertise.

By combining machine evaluation and manual evaluation, we can comprehensively and fairly assess the quality of each article. This dual evaluation method ensures the objectivity of the results while providing a deep understanding of the generated content. Ultimately, we calculate the total scores for the positive (A.1) and negative criteria (A.2) and derive the final composite score through the "positive minus negative" method. This scoring method ensures comprehensive and fair evaluation results, providing clear guidance for subsequent optimization and improvement.

This evaluation method and process demonstrate the strong potential of our system. Through a multi-layered architecture and complex interaction mechanisms, it provides an innovative and effective solution for future automated story generation.

In our experiments, the two major models that we tested were StabilityAI's StableLM 2 Zephyr 1.6B and Meta's Llama-2-7B both provided by hugging face. We utilized a variety of parameters to configure the Large Language Models (LLMs) for generating flash fiction stories. The key parameters employed in the model configuration are as follows:

- Max New Tokens: The maximum number of new tokens generated by the model was set to 3000.
- Temperature: The temperature parameter was set to 0.5.
- Top-K Sampling: The top-k parameter was set to 50. This sampling strategy restricts the model to consider only the top 50 most probable tokens at each step during text generation. This helps in maintaining a balance between creativity and coherence by avoiding less likely tokens that could disrupt the flow of the narrative.
- Top-P (Nucleus) Sampling: The top-p parameter was set to 0.95

These parameters were chosen to optimize the balance between creative diversity and narrative coherence in the generated stories. By fine-tuning these settings, we aimed to leverage the strengths of the LLMs in producing high-quality, engaging, and contextually rich flash fiction outputs. The results of our experiments indicated that these configurations played a crucial role in shaping the performance and effectiveness of the multi-agent communication pipelines used in this study.

5.3 Results

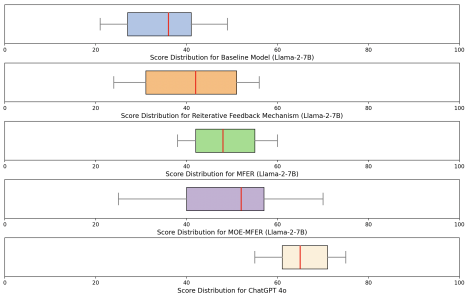

Figure 4: This box plot illustrates the distribution of scores for each of the different pipelines using the Llama-2-7B model compared to OpenAI GPT-4o.

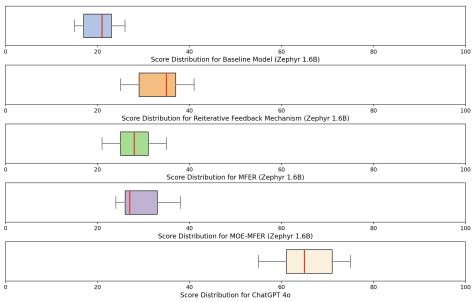

Figure 5: This box plot illustrates the distribution of scores for each of the different pipelines using the StabilityAI's Stable LM 2 Zephyr 1.6B compared to OpenAI GPT-4o.

We generally found that the pipelines dramatically increased the quality and score compared to the baseline of Llama-2-7B model with only the story prompt. For instance, for the set of results generated, the median for the Llama-2-7B baseline was 37.2, compared to 42.5 for the RFM Pipeline, 48.5 for the MFER pipeline, and finally 52.5 for the MOE-MFER pipeline. Important to notice is that the variation for the MOE-MFER pipeline was signficantly greater than the others, and sometimes performed at the level of the baseline. Out of these the only model that consistently seemed to outperform the baseline was the MFER.

We further attempted to apply these same pipelines to StabilityAI's Stable LM 2 Zephyr 1.6B. For the generated results, the corresponding medians were 22.1 for the baseline model, 37.6 for the RFM pipeline, 27.1 for the MFER pipeline and finally about 25.6 for the MOE-MFER pipeline. These values reflected the general trend that the values for Zephyr 1.6B a much smaller model were drastically lower. In this case, it seemed like the RFM model worked the best and seemed to be the only model that could consistently beat the baseline.

Overall, the quantitative results were largely within expectations, though some pipelines performed worse than anticipated on specific prompts. For example, we expected the MOE-MFER to consistently outperform the MFER, but the results showed dramatic variation that, in hindsight, are clearly possible. For the MOE-MFER pipeline, the results seem reasonable due to possibility that the selected words did not necessarily integrate well with the initial prompt and story. Further, including a dictionary

that was randomly chosen from introduced more avenues for creativity, but also likely resulted in large decreases in consistency due to variation differences. This suggests that while this approach has the potential to enhance model performance, the results were inconsistent and exhibited high variability. It indicates that there may be room for adjustments, such as changes to prompts and other factors, to address this issue. Furthermore, a non-trivial amount of variation was introduced purely by the metrics of evaluation as it is rather difficult, for both human evaluators and LLMs, to evaluate more subjective metrics such as creativity.

6 Analysis

Figure 6: Evaluations of different pipelines under Llama-2-7B model, with score obtained from Claude3 Opus judging (See Appendix for prompts)

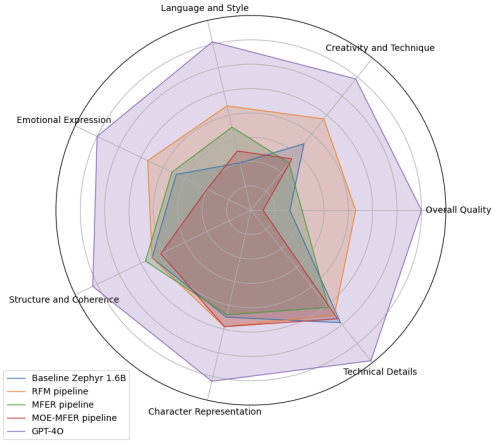

Figure 7: Evaluations of different pipelines under Zephyr 1.6B model, with score obtained from Claude3 Opus judging (See Appendix for prompts)

Your report should include *qualitative evaluation*. That is, try to understand your system (e.g., how it works, when it succeeds and when it fails) by inspecting key characteristics or outputs of your model.

6.1 RFM Analysis

Zephyr-1.6B When applied to this model, RFM shows a marked improvement over Baseline Zephyr in almost all areas. The reason why this pipeline seems to succeed with this model is because we do not rely on the model itself to act as a judge, instead using Gemini as a judge. As such, the model excels in creativity and technique, indicating innovative and effective execution. The pipeline's linguistic capabilities are quite strong, and it conveys emotions effectively.

Llama-2-7B This pipeline demonstrates slightly better overall quality than baseline Llama-2-7B, with notable improvements in creativity and technique. This suggests a higher degree of innovation in the iterative evaluation process where many evaluation dimensions are applied. Its language and style are slightly less polished, indicating minor areas for improvement in maintaining the linguistic qualities. The pipeline maintained a decent level of structure and coherence and emotional expression. Technical details are handled well, although slightly less than Llama-2-7B.

6.2 MFER Analysis

Zephyr-1.6B The model generally displayed lots of variability in its performance. From the text generated, it showed competent language and style, and effectively conveys emotions. Further, from scoring the narrative flow seems to be robust, suggesting a good structure and coherence. However, it lags in overall quality and creativity, indicating areas where it could improve. Its character representation and technical details are strong, but not the highest among the pipelines.

Llama-2-7B The MFER pipeline generally matches the RFM pipeline in overall quality, indicating a similar level of performance. Creativity and technique are on par with Llama-2-7B, suggesting great

innovative capabilities. This pipeline stands out in terms of of the emotional expression. However, the MFER pipeline falls slightly short in character representation, indicating some room for improvement in this area. Other criteria are maintained decently.

6.3 MOE-MFER Analysis

Zephyr-1.6B stands out in creativity and technique, language and style, and emotional expression. It, however, struggles significantly with maintaining a consistent narrative flow, which impacts its overall performance. Despite this, it excels in character representation and technical details, indicating a strong focus on these aspects.

Llama-2-7B The MOE-MFER pipeline shows a significant improvement in overall quality compared to the previous pipelines. It also demonstrates strong creativity and technique, indicating the successful approach in implementing a more robust pipeline which involves multiple layers. However, it shows some weaknesses in structure and coherence, suggesting slight issues in maintaining a consistent flow. Other criteria are maintained on a decent score.

7 Conclusion

In this project, we demonstrated three pipelines which leveraged multi-agents, role-playing, and iterative feedback for the writing of flash-fiction, or creative writing in general. These proposed pipelines were called Reiterative Feedback Mechanism (RFM), Modular Feedback-Enhanced Revision (MFER), and Mixture of Experts-Inspired Modular Feedback-Enhanced Revision (MOE-MFER) and each attempted applying and integrating different strategies as to improve ht estory output. RFM proved the most stable with tiny language models due to it's reliance on a much better trained and larger model, Gemini, to guide it's prompting as well as external resources in the form of great short stories. MFE proved to be generally the most reliable and was inspired in large part from many prior research on the subject area, further the mechanism by which it operated was consistent with expectations. Finally, MOE-MFER pipeline generally demonstrated clear trade-offs between creativity and consistency as it had highly variable performance. Even so, it illustrated great potential in creating far more interesting and well-written stories.

Our findings generally conclude that applying many experts and different forms of multi-agent collaboration can indeed enhance story-generation, in this case, for models that are open-source and even those that are quite small such as Zephyr-1.6B. The ability to write great flash fictions is likely transferable to other creative domains that people rely on LLMs for today. This is due to flash fiction representing a great proxy for other important creative tasks due to the need for conciseness and creative thinking while operating within the bounds of the prompt. In essence, when provided a prompt and little direction, these pipelines can sample greater areas of the creative space thus incorporating more interesting ideas. Future work could focus on refining these pipelines to reduce variability and further enhance the balance between creative innovation and narrative coherence. Additionally, it is important to investigate more sophisticated evaluation techniques and diverse datasets which could provide more insights into optimizing LLM-based story generation for broader applications.

8 Ethics Statement

The use of multi-agent large language models (LLMs) for creative writing might raise potential copyright issues. This is because moderate-sized LLMs such as Llama-2-7b may have already been trained on various existing artworks. Consequently, the LLM might inadvertently incorporate some features from these artworks, and the audience would not receive any details about which works were used during the generation process. For instance, these LLMs, in our flash fiction stories may "steal" or take ideas from existing authors. To mitigate this risk, one strategy could be integrating plagiarism detection into the system and enhancing the complexity of the prompts. This approach would help ensure that the system avoids inappropriately borrowing from previous works or outputting the result before it passes a certain threshold.

Another significant concern with the use of LLMs like these is the perpetuation of societal biases. These existing LLMs might have been trained on biased datasets, which can potentially carry out

stereotypical or offensive representations of certain groups of people. This occurs because the training data may reflect historical or societal biases that are then learned and perpetuated by the model. For example, if we feed these models stories that have people in traditional roles of power like men as the heroes in each story, we may unintentionally proliferate these biases to people who read the stories generated by these large language models. One potential mitigation technique to address this issue is to implement restrictions on the prompts given to the model. By doing so, the model can be guided to avoid writing on specific sensitive topics. Additionally, this strategy encourages the model to figure out alternative creative ways to generate stories, thus significantly lowering the risk of harming certain groups or perpetuating existing biases. This approach not only aims to prevent the generation of offensive content but also promotes a more inclusive and considerate use of technology in creative writing.

References

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

[2] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. 2023.

[3] Bin Lei. Macm: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. 2024.

[4] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. 2024.

[5] Yoichi Ishibashi and Yoshimasa Nishimura. Self-organized agents: A llm multi-agent framework toward ultra large-scale code generation and optimization. 2024.

[6] Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. 2023.

[7] Piotr Wojciech Mirowski, Juliette Love, Kory W. Mathewson, and Shakir Mohamed. A robot walks into a bar: Can language models serve as creativity support tools for comedy? an evaluation of llms' humour alignment with comedians. 2024.

[8] Filippo Chiarello, Vito Giordano, Irene Spada, Simone Barandoni, and Gualtiero Fantoni. Future applications of generative large language models: A data-driven case study on chatgpt. *Technovation*, 133:103002, May 2024.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. 2022.

[10] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[11] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[12] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. 2021.

[13] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint*, 2023.

[14] Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xingxu Xie. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint*, 2023.

[15] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity, 2024.

[16] Carlos Gómez-Rodríguez and Paul Williams. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. pages 14504–14528, December 2023.

[17] Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2024.

[18] Reedsy. The best flash fiction of 2021, 2021. Accessed: 2024-06-07.

[19] Globe Soup. 100 Awesome Flash Fiction Prompts - Plus Bonus Prompts! — GLOBE SOUP — globesoup.net. `https://www.globesoup.net/writing-blog/100-awesome-flash-fiction-prompts`. [Accessed 24-05-2024].

A Appendix (optional)

A.1 Positive Criteria (each criterion with a maximum score of 10 points)

- **Overall Quality**:
  - Overall Impression
  - Overall Reader Engagement
  - Compelling Ending
- **Creativity and Technique**:
  - Clever/Witty
  - Gripping
  - Effective Use of Tropes
  - Terse Stream-of-Consciousness Style
- **Language and Style**:
  - Sentences Flow Naturally
  - Well-earned Lightness or Darkness
  - Elegant Prose
  - Imagery and Descriptive Quality
  - Consistent Voice/Tone of Writing
- **Emotional Expression**:
  - Emotionally Complex
  - Emotionally Engaging
- **Structure and Coherence**:
  - Coherent
  - Appropriate Length
  - Dialogue is Naturally Integrated
  - Unreliable Narration
- **Character Representation**:
  - Believable Characters
  - Believable Character Actions
  - Adherence to Character Bios
- **Technical Details**:
  - Correct Spelling & Grammar
  - Adherence to Instructions
  - Diary Entries Feel Natural

A.2 Negative Criteria (higher scores indicate poorer performance)

- **Dialogue Quality**:
  - Repetitive Tit-for-Tat Dialogue
  - Stilted dialogue
- **Language Issues**:
  - Clunky Asides and Interruptive Sentence Structures
  - Amateurish Descriptives
  - Profundity Over-reach
- **Moral and Emotional Issues**:
  - Simplistic Moralizing
  - Shallow Optimism
  - Melodramatic
- **Plot and Character Issues**:
  - Unearned Transformations
  - Incongruent Ending Positivity
  - Characters are Too Good
  - Shallow Resolution
- **Creativity and Inspiration Issues**:
  - Trite
  - Overwrought
  - Amateurish
  - Contrived
  - Uninspiring

- **Romantic Emotional Changes**: How do romantic relationships between characters develop? Are there new romantic subplots?
- **Friendship Emotional Changes**: How does friendship impact the plot? Are there betrayals or new friendships?
- **Conflict**: What are the main conflicts in the story? How are these conflicts intensified or resolved?
- **Character Growth**: How do characters grow and change throughout the story? What challenges and successes do they face?
- **Secret Revelation**: Are there hidden secrets revealed in the plot? How do these secrets impact characters and story development?
- **Mission Goals**: What are the main goals of the characters? How do they strive to achieve these goals?
- **Humorous Misunderstandings**: Are there funny or misunderstanding subplots that add humor to the story?
- **Social Reflection**: How does the story reflect social issues or realities? How are these elements integrated into the plot?
- **Internal and External Conflicts**: What are the internal conflicts and external challenges of the characters? How are these conflicts managed?

If you wish, you can include an appendix, which should be part of the main PDF, and does not count towards the 6-8 page limit. Appendices can be useful to supply extra details, examples, figures, results, visualizations, etc. that you couldn't fit into the main paper. However, your grader *does not* have to read your appendix, and you should assume that you will be graded based on the content of the main part of your paper only.