HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 4 (2024)

Efficient Visual Prompt Engineering for Creative Story Writing

Felix Deng

Redmond High School, USA

ABSTRACT

Large Language Models (LLMs) [1] are extensively utilized for generating stories, showcasing their ability to handle complex, creative tasks. To begin the process of story generation, an initial textual prompt is required. The prompt is iteratively refined such that the discrepancy between the user's expectations and the story generated from the prompt is minimized. Each iteration is a time-consuming process; the user needs to read and analyze the story in order to refine the prompt. A key insight from cognitive research suggests that analyzing visual data is 60,000 times faster than textual analysis [2]. This paper proposes visual prompt engineering for story generation wherein textual prompts are transformed into images using a diffusion model [3], then refined based on the discrepancy between the user's expectations and the generated image. This refined prompt is then used to generate a story. The entire process is repeated until the user is satisfied with the story. This method leverages the relative speed of image processing to enhance the quality of text generation per iteration. Experiments show that for the same number of iterations, stories generated by visual prompt engineering outperformed those generated by text-based prompts in terms of story quality.

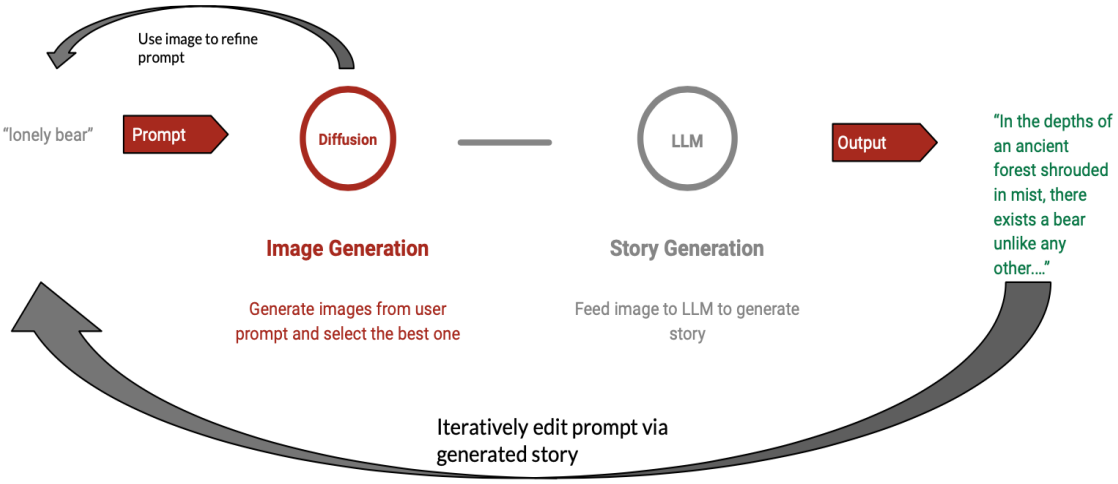


Figure 1. Visual prompt engineering.

Introduction

Generative AI has revolutionized various fields by enabling the creation of content through large language models. One of the critical advancements in this area is the use of prompt engineering, which involves optimizing inputs to guide the behavior and outputs of language models (LLMs). This technique helps in steering the models to generate text that is more relevant to the prompt, enhancing their usability in creative and practical applications.

Despite these advancements, leveraging LLMs to create high-quality stories presents significant challenges. The iterative process of text generation is inherently time-consuming, requiring careful tuning of prompts and multiple rounds of generation to achieve the desired outcome. Furthermore, reading and evaluating the generated text further increases the time spent, making the process labor-intensive.

Related Work

Iterative Prompting

Iterative prompting [4] involves refining the prompts given to a Large Language Model (LLM) through multiple rounds of feedback and adjustment. This process is akin to an interactive dialogue with the model, where initial outputs are reviewed, and the prompts are tweaked to steer the model towards more accurate or creative outputs. Iterative prompting can be manual, involving human oversight, or automated using feedback loops within the AI system.

The exact process of iterative prompting is as follow:

1. User provides a broad initial prompt, such as "generate a story about a bear."
2. The response from the model is then analyzed to assess its relevance and quality. Based on the user's expectations, the prompt is refined to be more specific, such as "generate a story about a brown bear."
3. This cycle of generating, analyzing, and refining is repeated until the desired output is achieved. This iterative approach helps in honing the model's responses to be more aligned with the user's expectations.

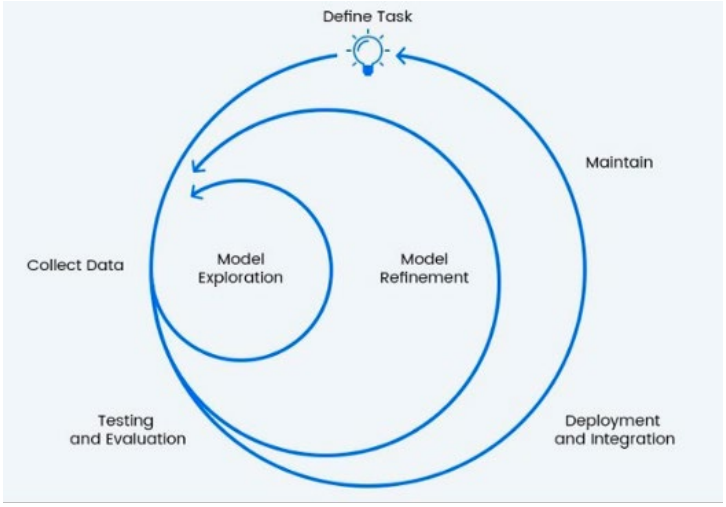


Figure 2. Iterative prompting. [5]

Despite its effectiveness, iterative prompting has several limitations. The process is time-consuming, as it requires the user to read and evaluate the results after each iteration. It also requires the user to optimize the original prompt for each iteration, another lengthy process. This need for continuous oversight and optimization can be labor-intensive, especially for complex tasks or long outputs.

Auto Prompting

Auto prompting [6] refers to the automated generation of input prompts for Large Language Models (LLMs). This technique involves the use of algorithms or smaller, specialized models to craft prompts that are optimized to elicit

the best possible response from an LLM. The goal is to reduce human labor in the prompt engineering process and to increase the efficiency and consistency of the outputs. Auto Prompting can be particularly useful in applications like chatbots, content generation, and other interactive AI systems where dynamic and contextually appropriate responses are crucial.

The exact process of auto prompting is as follows:

1. The user provides an initial prompt and relevant labeled training data to the system.
2. A masked language model (MLM) is trained based on the data.
3. Based on the data, prompts are generated and refined based on the output of the MLM. This process is repeated until high-quality responses are generated.
4. After the prompt is optimized, it is returned to the user.

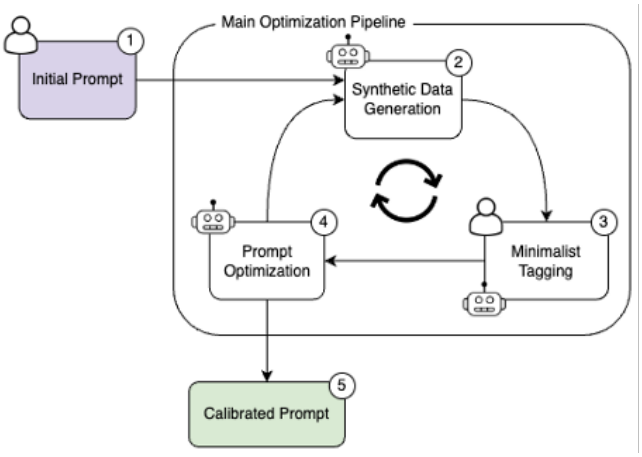


Figure 3. Auto prompting. [7]

Although auto prompting significantly boosts prompt quality, one of the primary limitations of Auto Prompting is the time required for optimization. The process of refining a prompt can take several minutes, which may not be suitable for applications requiring immediate responses. Additionally, the computational resources required for this optimization translate into costs, with each optimization cycle potentially costing around $1 [8].

Solution

To address these challenges, a solution is proposed in the paper where image generation is integrated into the creative process. Diffusion models, which convert textual descriptions into visual content, offer a promising alternative. Analyzing images is 60,000 times faster than analyzing text [9] due to the innate human ability to process visual information. This assumption is supported by cognitive studies that show faster comprehension and retention of visual data compared to textual data.

To assess the viability of integrating image generation into story creation, it is crucial to demonstrate that this approach is more efficient than the baseline, iterative prompting.

How does the incorporation of visual imagery in prompt generation affect the efficiency of the iteration process?

Experiment Design

To evaluate the effectiveness of image-based prompt engineering, we undertook a comprehensive study involving the comparison of an iterative prompting Method And image-based prompt engineering to test the efficiency of our approach.

Methods

For this experiment, two methods, methods A and B, were used. Method A is the baseline model, where text is directly sent to a LLM to produce a story. Method B uses image generation, as text is converted to an image through DALL-E 2 and then converted into text using an image-to-text model.

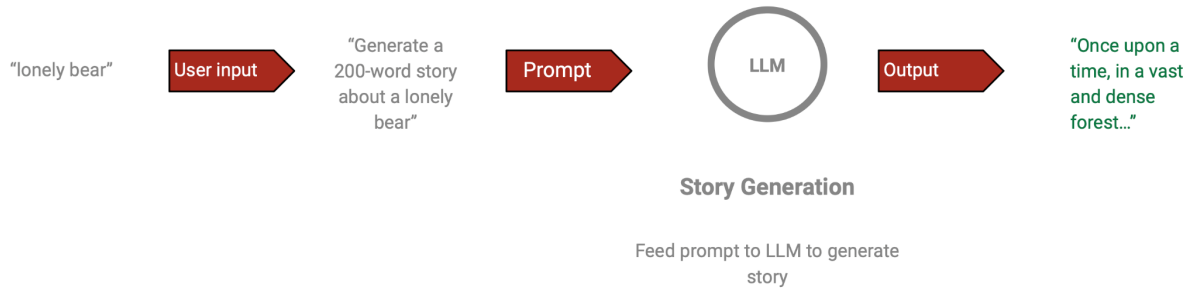


Figure 3. Diagram of Method A.

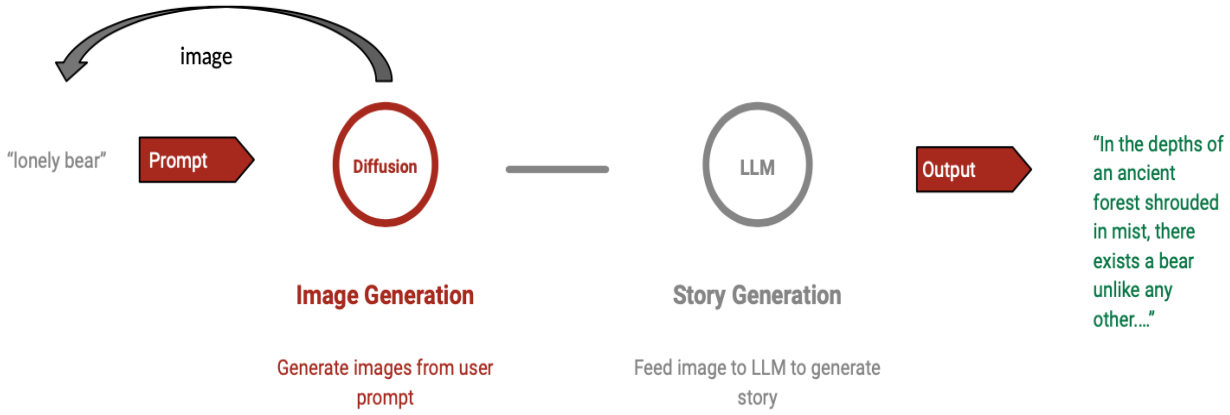


Figure 4. Diagram of Method B.

Time

One of the main concerns with visual prompting is the increase in time per iteration compared to the baseline. However, this time increase was found to be negligible.

The amount of time per iteration for Method A was found to be $T_A = T_{read} + T_{generate}$

The amount of time per iteration for Method B was found to be $T_B = T_{read} + k * (T_{image} + T_{analyze})$

T_{read} is the amount of time needed to read and analyze a 500-word story, roughly 21 minutes. This is because of the many small components of this task:

- Reading the story: The average human reading speed is 200-300 words per minute [10], so one full reading will take 2.5 minutes. For further comprehension, another readthrough will be necessary, so 5 minutes will be spent on reading in total.
- Analyzing the story: An editor can edit roughly 500 words in 16 minutes. [11]

$T_{generate}$ is the amount of time needed to generate a 500-words story, roughly 15 seconds. I timed GPT-4 50 times while it generated stories to get the average amount of time.

k is the amount of iterations to fully refine a prompt based on generated images, on average around 5 iterations.

T_{image} is the amount of time to generate an image based on a prompt, also roughly 15 seconds. I timed my diffusion model 25 times while it generated stories to get the average amount of time.

$T_{analyze}$ is the amount of time for the user to analyze an image, an amount of time which is negligible considering that analyzing images is much faster than analyzing text.

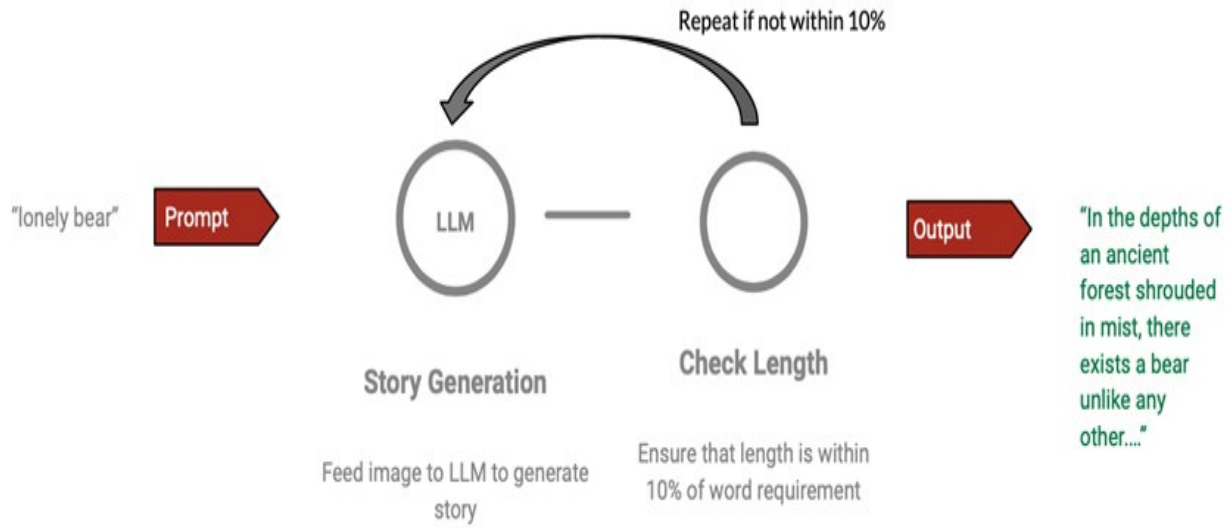
Thus, Method B takes roughly 75 seconds longer than Method A per iteration. This slight time increase can be ignored, as the story analysis process takes 21 minutes on average.

Generating Story Subjects

For each genre, 200 varying story subjects were generated. A story subject refers to the main character or central theme of a story, such as "a lonely bear." These subjects were used as prompts to guide the LLMs in story creation.

Iterative Story Generation Process

1. Initial Prompting: The story subject is fed into Method A and Method B.
2. Word Count Adjustment: Each generated story should be roughly 500 words. To generate these stories, the iterative prompting process was employed, where:



3. Refinement: Users refine the story subject of both Method A and Method B by reducing discrepancies between their expectations and either the generated story from Method A or the generated image from Method B.

4. Iteration: Steps 1-3 are repeated either 5, 10, or 20 times to ensure quality. In the end, the prompts from Method A and Method B are converted into stories for grading.

Assignment and Grading of Stories

1. Random Assignment: 30 prompts and 60 stories, 30 generated by Method A and 30 generated by Method B, were randomly selected. All 60 stories were each assigned to GPT-4, which graded the stories based on the rubric below.

2. Grading Criteria: Each story was graded on a scale of 1-10 based on five key dimensions:
   - Imagery/Setting: The vividness and clarity of the story's setting and imagery.
   - Character Development: The depth and believability of the characters.
   - Mood/Tone: The emotional tone and atmosphere of the story.
   - User Engagement: How captivating and engaging the story is.
   - Originality: The uniqueness and creativity of the story.

3. Average Scoring: For each story, the average score across the five dimensions was calculated to determine its overall quality.

This structured approach allowed us to systematically evaluate the quality of stories generated by LLMs and the influence of visual imagery on narrative quality. By using an objective grading system (GPT-4) and rigorous evaluation criteria, we ensured a robust assessment of the stories' effectiveness in various key areas of storytelling. This methodology provides valuable insights into the potential benefits and limitations of integrating image generation into LLM-driven story creation.

Results

Table 1. Average results of methods A and B based on number of iterations, given a 500 word requirement.

Iterations	Method A	Method B
1 time	7.4	7.8
2 times	8.4	8.6
5 times	8.7	8.8

The incorporation of image generation significantly enhances the efficiency of the story creation process. By providing visual context, the diffusion model guides the LLM more effectively, reducing the number of iterations required to achieve the desired output. This reduction in iterative cycles translates directly into time savings, making the process faster and more efficient without compromising the quality of the generated stories.

Note that as iterations increase, the increase in quality decreases sharply. For Method A, quality increases by 1.0, then 0.3. This is because after a certain number of iterations, both models will produce a maximum-quality output. Thus, as models approach this limit, the rate of change of the quality will decrease.

Thus, integrating image generation offers a more efficient alternative, as it produces higher-quality results in less iterations.

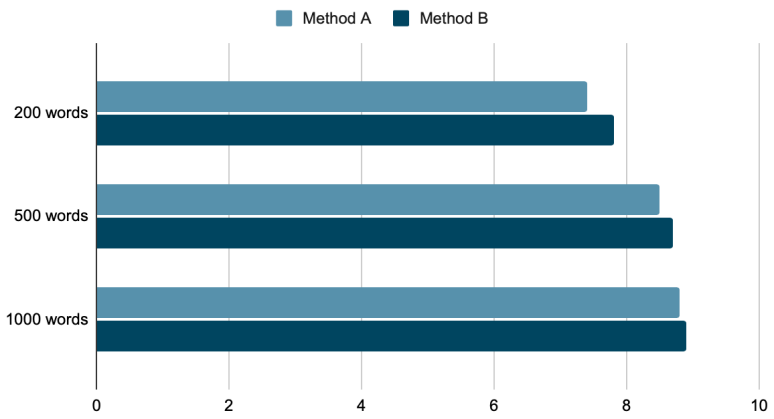


Figure 4. Average results of methods A and B based on word requirement, given 2 iterations.

As shown in the above table, method B consistently outperforms method A regardless of the word count. However, as the word requirement increases, the quality of stories generated by both methods approaches a maximum value. The difference in quality between method B and method A narrows over time, decreasing from an initial difference of 0.5 to a minimal difference of 0.1.

This trend indicates that while method B initially has a distinct advantage, the benefits diminish as the length of the story increases. This suggests a potential plateau effect where both methods achieve near-optimal quality for longer stories.

The diminishing difference in quality could imply that as more content is generated, the methods converge in their ability to produce high-quality narratives. This convergence might be due to the increased opportunities for both methods to refine and enhance the story through iterative processes, ultimately leading to a more polished final product.

Sample Stories

Both stories are roughly 200 words, generated from the prompt “a lonely bear”.

Method A’s story is shown below:

Once upon a time, in a vast and dense forest nestled between towering mountains, there lived a bear named Oliver. Oliver was not like the other bears in the forest. While they roamed in groups, playing and hunting together, Oliver preferred the solace of his own company. He was a solitary soul, finding solace in the quietness of the woods. From a young age, Oliver had felt different. While the other bear cubs wrestled and played, he often found himself wandering off, drawn to the tranquility of the forest. His parents worried about him, urging him to join the other cubs in their activities, but Oliver couldn't shake off the feeling of isolation that seemed to permeate his very being. Oliver's solitary nature became more pronounced as he grew older. He found joy in exploring the hidden paths of the forest, discovering its secrets and observing its inhabitants from a distance. His keen senses and quiet demeanor earned him a reputation among the animals as a thoughtful and observant bear. While his fellow bears engaged in boisterous activities, Oliver would often be found near a babbling brook or under the shade of ancient trees, lost in thought.

Method B’s story is shown below:

In the depths of an ancient forest shrouded in mist, there exists a bear unlike any other. Named Baran, he was born with an extraordinary coat, a soft golden-brown hue that shimmered when the rare sunlight filtered through the dense

canopy. Despite his magnificent appearance, Baran was a solitary figure, his heart weighed heavy with a deep longing for companionship. Baran's mother had often told him stories of the world beyond their forest, of lands filled with bears of various kinds and tales of grand adventures. But those were just stories. In reality, Baran's forest was remote, and he remained the last of his kind in these parts, a gentle giant whose voice had not found an echo for years. Baran's days were filled with quiet contemplation and exploration. He would roam the moss-covered trails, pausing to listen to the whispers of the ancient trees and the melodies of the forest creatures. Despite his solitude, he harbored a flicker of hope that one day he might encounter another bear like himself. His golden-brown coat stood out against the lush greens and earthy browns of the forest, a beacon of his uniqueness in the vast wilderness. Yet, with each passing season, his longing for companionship deepened, echoing through the silent woods.

Discussion

The integration of visual imagery into narrative generation using Large Language Models (LLMs) presents a promising avenue for enhancing storytelling efficiency and quality. The findings of our study suggest that using visual prompts can streamline the story generation process, providing a faster alternative to traditional auto prompting techniques. By incorporating images, the iterative process of prompt refinement and story generation can be expedited, making it a more practical approach for applications requiring dynamic content creation.

Conclusion

Overall, incorporating a diffusion model to enhance prompt engineering, particularly in creative story writing, has proven to be more efficient. By integrating visual prompts, the iterative process of refining and generating stories is streamlined, reducing the time and effort required compared to traditional methods. The use of images quickly establishes context, guiding LLMs to produce more relevant and engaging narratives.

Despite its success, several limitations need to be addressed for broader application. The current focus on single-subject prompts and restriction to story generation highlight the need for further exploration into its utility across diverse text generation domains. Enhancing the system's ability to handle more complex and multifaceted prompts could lead to richer and more varied narratives, expanding its potential applications.

While the diffusion model for enhancing prompt engineering in creative story writing has shown significant promise, addressing its limitations through targeted research and development will be crucial for unlocking its full potential across various text generation tasks.

Applications

One of the primary applications of incorporating visual imagery into prompt engineering is in story generation, where visual prompts can quickly establish context and stimulate creative outputs from Large Language Models (LLMs). This technique can be particularly beneficial for developing authors, game designers, and other creative professionals who rely on high-quality storytelling. By providing a visual context, these professionals can generate more vivid and immersive narratives, enhancing the reader's or player's experience.

Visual imagery can serve as an immediate source of inspiration, helping authors to create detailed and imaginative settings, complex characters, and engaging plotlines. For instance, an image of a medieval castle surrounded by dark forests can evoke a multitude of narrative possibilities, from epic battles to mystical adventures. This approach can help authors overcome creative blocks and produce content that is both rich in detail and emotionally engaging.

In the realm of video game design, visual prompts can assist in the development of storylines and character backstories, making the game world more immersive and believable. Game designers can use visual imagery to create

detailed environments and intricate plots that enhance the player's experience. For example, an image of an abandoned spaceship can inspire a thrilling sci-fi adventure game, complete with alien encounters and mysterious technologies.

Visual imagery is especially useful for mitigating writer's block. When writers are unclear on what they want to write about, visual prompts can serve as effective guidelines to help guide their thoughts towards an idea for a story. The image acts as a catalyst, sparking imagination and providing a concrete starting point from which writers can develop their narratives. This can be particularly beneficial for novice writers who may struggle with generating ideas or for experienced writers who face occasional creative slumps.

Overall, the integration of visual imagery into prompt engineering offers a versatile tool that can enhance creativity and productivity across various fields. By providing a rich source of inspiration and a concrete framework for narrative development, visual prompts can help individuals and professionals produce high-quality, engaging content more efficiently.

Limitations

The experiment has certain limitations which may have skewed the results of the experiment.

Firstly, only 200 story subjects were generated. This results in a small sample size, which may impact the results of the experiment. In the future, 1000 or more story subjects should be generated for future replications of this experiment.

Additionally, stories were graded using a specific rubric. While the rubric contains important features which every story should have, it fails to include elements such as cohesion and readability. Additional quantitative metrics, such as the Automated Readability Index (ARI) [12] should be used in future experiments.

Stories were also graded using GPT-4, which interpreted the rubric. GPT-4 may not evaluate stories in the way which an average human does, which may lead to inaccurate results. To remedy this, a large sample of human graders or a fine-tuned variant of GPT-4 should be used for future experiments.

Future Research

Future research should explore the potential of integrating visual prompts in various text generation tasks beyond storytelling. For instance, visual prompts could be used to enhance the creation of informative articles by providing a visual context that helps the LLM generate more accurate and detailed content. Similarly, in persuasive writing, images can evoke emotions or highlight key points, potentially making arguments more compelling. Instructional content could also benefit from visual aids, as images can clarify complex concepts and make instructions easier to follow. Investigating these applications could broaden the utility of visual prompts in diverse fields.

Another promising area for future research is the development of techniques to handle more complex and multifaceted prompts. Currently, the focus has been on simple, single-subject prompts, but more intricate and layered prompts could lead to richer and more diverse narratives. For example, a prompt combining multiple characters, settings, and plot elements could produce a more elaborate and engaging story. Developing methods to effectively manage and refine these complex prompts will enhance the creative capabilities of LLMs and allow for the generation of more sophisticated and nuanced content.

Ensuring that the generated text accurately reflects the provided visual prompts is crucial for the success of this approach. Future research should focus on improving the alignment between images and the resulting narratives. This could involve developing advanced algorithms that better interpret visual cues and translate them into coherent and contextually appropriate text. Enhancing image-to-text alignment will reduce inconsistencies and inaccuracies in the generated stories, resulting in higher-quality content that better meets user expectations.

Endnotes

1: OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024, March 4). *GPT-4 technical report*. arXiv.org. https://arxiv.org/abs/2303.08774. https://doi.org/10.48550/arXiv.2303.08774

2: Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022, May 23). *Photorealistic text-to-image diffusion models with Deep Language understanding*. arXiv.org. https://arxiv.org/abs/2205.11487. https://doi.org/10.48550/arXiv.2205.11487

3: Pamplona, F. (2023, September 22). *Column - the power of visuals*. MedTech Intelligence. https://medtechintelligence.com/column/the-power-of-visuals/#:~:text=The%20human%20brain%20processes%20images,also%20more%20memorable%20than%20text

4: Hosni, Y. (2024, February 20). *Prompt engineering best practices: Iterative prompt development*. Medium. https://pub.towardsai.net/prompt-engineering-best-practices-iterative-prompt-development-22759b309919?gi=e0a65467b4ea

5: Hosni, Y. (2024, February 20). *Prompt engineering best practices: Iterative prompt development*. Medium. https://pub.towardsai.net/prompt-engineering-best-practices-iterative-prompt-development-22759b309919?gi=e0a65467b4ea

6: Levi, E., Brosh, E., & Friedmann, M. (2024, February 5). *Intent-based prompt calibration: Enhancing prompt optimization with synthetic boundary cases*. arXiv.org. https://arxiv.org/abs/2402.03099. https://doi.org/10.48550/arXiv.2402.03099

7: Levi, E., Brosh, E., & Friedmann, M. (2024, February 5). *Intent-based prompt calibration: Enhancing prompt optimization with synthetic boundary cases*. arXiv.org. https://arxiv.org/abs/2402.03099. https://doi.org/10.48550/arXiv.2402.03099

8: Levi, E., Brosh, E., & Friedmann, M. (2024, February 5). *Intent-based prompt calibration: Enhancing prompt optimization with synthetic boundary cases*. arXiv.org. https://arxiv.org/abs/2402.03099. https://doi.org/10.48550/arXiv.2402.03099

9: Pamplona, F. (2023, September 22). *Column - the power of visuals*. MedTech Intelligence. https://medtechintelligence.com/column/the-power-of-visuals/#:~:text=The%20human%20brain%20processes%20images,also%20more%20memorable%20than%20text

10: Carver, R. P. (1990). "Reading Rate: A Review of Research and Theory." Academic Press.

11: Babcock, L. (2024, August 6). *How long proofreading takes (a complete guide with tables)*. Om Proofreading. https://omproofreading.com/how-long-proofreading-takes/

12: Kincaid, J. P., & Delionbach, L. J. (1973). Validation of the Automated Readability Index: A Follow-Up. Human Factors, 15(1), 17-20. https://doi.org/10.1177/001872087301500103

References

Levi, E., Brosh, E., & Friedmann, M. (2024, February 5). *Intent-based prompt calibration: Enhancing prompt optimization with synthetic boundary cases*. arXiv.org. https://arxiv.org/abs/2402.03099. https://doi.org/10.48550/arXiv.2402.03099

Singh, C., Morris, J. X., Aneja, J., Rush, A. M., & Gao, J. (2023, January 26). *Explaining patterns in data with language models via interpretable autoprompting*. arXiv.org. https://arxiv.org/abs/2210.01848. https://doi.org/10.48550/arXiv.2210.01848

Wang, B., Deng, X., & Sun, H. (2022, October 23). *Iteratively prompt pre-trained language models for chain of thought*. arXiv.org. https://arxiv.org/abs/2203.08383. https://doi.org/10.48550/arXiv.2203.08383

Pamplona, F. (2023, September 22). *Column - the power of visuals*. MedTech Intelligence. https://medtechintelligence.com/column/the-power-of-visuals/#:~:text=The%20human%20brain%20processes%20images,also%20more%20memorable%20than%20text

Hosni, Y. (2024, February 20). *Prompt engineering best practices: Iterative prompt development*. Medium. https://pub.towardsai.net/prompt-engineering-best-practices-iterative-prompt-development-22759b309919?gi=e0a65467b4ea

Tuscher, M., Schmidt, J. (2022, October 16). *Processing speed and comprehensibility of visualizations*. VRVis Forschungs-GmbH. https://www.vrvis.at/publications/pdfs/PB-VRVis-2022-016.pdf

Kincaid, J. P., & Delionbach, L. J. (1973). Validation of the Automated Readability Index: A Follow-Up. Human Factors, 15(1), 17-20. https://doi.org/10.1177/001872087301500103

Carver, R. P. (1990). "Reading Rate: A Review of Research and Theory." Academic Press.

Babcock, L. (2024, August 6). *How long proofreading takes (a complete guide with tables)*. Om Proofreading. https://omproofreading.com/how-long-proofreading-takes/