NLP Report - 2

Group - 6

Parth Sudan - 2022A7PS0177P
Dhruv Shrimali - 2021A7PS0008P

Inspiration for Using "GenAssist: Making Image Generation Accessible"

The decision to use GenAssist: Making Image Generation Accessible as a foundational paper stemmed from its alignment with the core objectives of our project. The original problem statement emphasizes overcoming language barriers in a multilingual and diverse society by leveraging visual communication to make welfare schemes accessible. Similarly, GenAssist focuses on democratizing image generation by bridging the gap between complex AI models and non-expert users through intuitive systems.
The parallels between the paper's goal and our project's aims are clear: both prioritize accessibility and user-centric design in the context of visual content creation. The methodologies outlined in GenAssist - such as simplifying the process of prompt creation and ensuring generated images align with intended purposes - are particularly relevant to our assignment objectives. These include generating detailed image prompts from textual descriptions, maintaining entity consistency, and evaluating coherence in the generated images.

Improvements We Are Hoping For

While GenAssist provides a solid foundation for making image generation accessible, our project builds on its principles with a focus on addressing specific challenges encountered during our initial implementation, particularly in the context of welfare scheme communication in India.

1. ## Localized Adaptation and Cultural Sensitivity:

   Unlike GenAssist, which uses a more general approach, our project focuses on welfare schemes in India. As we observed in our initial implementation, there is a need to tailor prompts to reflect the diversity of Indian culture, regions, and languages. This includes generating culturally relevant imagery, such as traditional attire, rural and urban settings, and other region-specific features. Our goal is to generate images that resonate with the specific contexts in which welfare schemes will be implemented, ensuring that visual depictions are both meaningful and relatable.

2. ## Legibility and Text Clarity in Indian Languages:

   One of the significant limitations we encountered with DALL-E-3 was its inability to generate legible text in various languages, including English, Hindi, and other Indian languages. In several cases, the model produced distorted or garbled text, which severely impacted the clarity of images where text was a key element (e.g., scheme names or important information). A major improvement we are hoping to achieve is enhancing the model's ability to generate clear and readable text in these languages,

ensuring that textual details remain legible, coherent, and useful for the intended audience, or removing it completely.

3. Improved Consistency in Art Style:

During our first round of testing, we noticed inconsistencies in the art style of the generated images. This lack of consistency detracted from the overall visual cohesion of the welfare posters. Our aim in the next stage is to refine the prompts and implement strategies that ensure a more consistent and uniform art style across all images generated for a given scheme. This will improve the overall visual coherence and make the images more professional and recognizable.

4. Enhanced Entity Consistency Across Multiple Prompts:

While GenAssist provides a general approach to prompt generation, our project goes further by focusing on maintaining entity consistency across different prompts. In our initial tests, we observed inconsistencies in how characters, objects, and settings were depicted across different prompts for the same welfare scheme. We plan to implement more sophisticated techniques for ensuring that key entities (such as characters or specific objects) are consistently represented across all prompts and iterations, reinforcing the coherence and clarity of the generated images.

5. Objective Evaluation Based on Visual Quality and Relevance:

In line with GenAssist's focus on accessibility, our project takes a more detailed approach to the evaluation of generated images. Specifically, we plan to develop systematic criteria for evaluating image relevance, coherence, and adherence to the original prompt. This will include evaluating how well the generated images align with the welfare scheme's goals and messages. By refining our evaluation process, we aim to ensure that the generated posters are not only visually appealing but also effectively communicate the intended information.

6. Automation Pipeline for Efficiency and Scalability:

Inspired by GenAssist’s approach to simplifying prompt generation, we aim to develop an automated pipeline that covers the entire workflow - from information extraction from welfare scheme documents to prompt generation and image refinement. In particular, the pipeline will include an iterative refinement process that can automatically adjust and improve image prompts based on the feedback from visual quality assessments and entity consistency checks. This will allow for a more efficient and scalable approach to producing large-scale, customized visual content for welfare schemes.

7. Impact on Public Welfare:

   Our project goes beyond the theoretical framework of GenAssist to have a real-world impact. By improving the legibility, visual clarity, and consistency of generated images, and ensuring cultural relevance in the design of welfare scheme posters, we aim to bridge the information gap for underserved communities. The ultimate goal is to create visually engaging and accessible posters that effectively communicate key welfare information, ensuring that vital schemes reach people, irrespective of their language or literacy levels.

Implementation

Step 1: Information Extraction (same as before)

The first step involves processing welfare scheme documents to extract critical details. These documents typically contain information about the scheme's goals, target audience, location, and other specific elements like beneficiaries, resources, and timelines.
Techniques are applied to extract key entities such as names of people, organizations, locations, and dates, as well as important phrases that describe the scheme's objectives. For example, extracting "education for girls in rural areas" would lead to the creation of a more targeted visual representation.

Step 2: Image Prompt Generation (same as before)

After extracting the relevant information, the next step is to convert this data into descriptive image prompts. These prompts are designed with enough detail to guide text-to-image model (dall-e-3) in generating appropriate visuals. The same prompt as submission-1 was used.

Step 2.5: Generating Character Descriptions and Art Style (introduced later)

Before generating the images, we introduce an important step: generating detailed character descriptions and specifying the art style for the welfare scheme visuals. This step ensures that the characters are represented consistently and in line with the intended aesthetic for the posters.
In this step, the LLM is tasked with generating descriptions for 2-3 main characters, along with the art style to be used in the images. The descriptions focus on the outward appearance of the characters, ensuring they are clearly defined in terms of their physical features, without specifying the background or facial expressions. The art style is also specified in broad terms, such as realistic, manga, or cartoon, based on the nature of the scheme.

These generated descriptions and art style details are appended to the image generation prompts, providing clear guidance for the text-to-image models. This helps maintain consistency across character representations while leaving the background and facial expressions to be determined by the context and scene later in the process.

Step 3: Initial Image Generation (same as before)

The structured prompts are then inputted into a text-to-image generation model (dall-e-3). This model generates an initial set of images based on the provided prompts. During this stage, several parameters are crucial:

- Style Consistency: Ensuring the generated images maintain a consistent artistic style.
- Entity Representation: Accurately portraying key entities mentioned in the prompt (e.g., people, locations, resources).

These generated images are stored for evaluation in subsequent steps.

Step 4: Generating IBQ and PVQ

Along with generating images, prompts are structured to facilitate efficient evaluation using Image based Questions (IBQ) and Prompt Verification Questions (PVQ):

1. Prompt Verification Questions (PVQ):

   PVQs are designed to verify whether the generated image accurately matches the description provided in the prompt. These questions focus on confirming whether key elements, such as characters, actions, objects, and settings mentioned in the prompt, are represented correctly in the image. For instance, a PVQ might ask:
   - "Does the image accurately depict the rural setting mentioned in the prompt?"
   - "Are the key figures or objects from the prompt, such as a classroom or children, visible and correctly represented?"
     The goal is to ensure the alignment of the image with the intent of the prompt and to check if the generated image includes all the necessary components.
2. Image-Based Questions (IBQ):

   IBQs encourage a deeper, more detailed observation of the image. These questions focus on aspects that may not have been explicitly stated in the prompt but are important for understanding the image in its entirety. For example, IBQs could ask about:
   - "What is the significance of the background elements in the image?"
   - "Can you identify any cultural or contextual details that add meaning to the scene?"
   - "How are the characters' clothing or posture reflecting the actions or environment?"
     These questions are meant to help assess the subtle details, emotional tone, and narrative conveyed through the visual content, going beyond the explicit instructions in the prompt.

Step 5: Evaluation of Generated Images

Once the initial images are generated, they are evaluated based on their alignment with the original prompt. The evaluation process includes several key metrics:

- Relevance: How accurately the image represents the extracted information from the prompt.
- Coherence: The logical flow and harmony within the image, ensuring all elements work together effectively.
- Text Legibility: Ensuring that any textual content in the image is clear and readable.

Gemini-flash-002 was provided with the generated image and asked to answer both Prompt Verification Questions (PVQ) and Image-Based Questions (IBQ). These questions were designed to:

- Verify whether the key elements mentioned in the prompt, such as characters, actions, objects, and settings, are accurately reflected in the image (PVQs).
- Encourage a detailed analysis of the visual elements in the image, such as specific objects, attire, and background elements that may not be explicitly mentioned in the prompt (IBQs).

This evaluation process helps ensure that the generated image aligns well with the intended vision, highlighting areas of strength and areas for improvement in the image quality and accuracy.

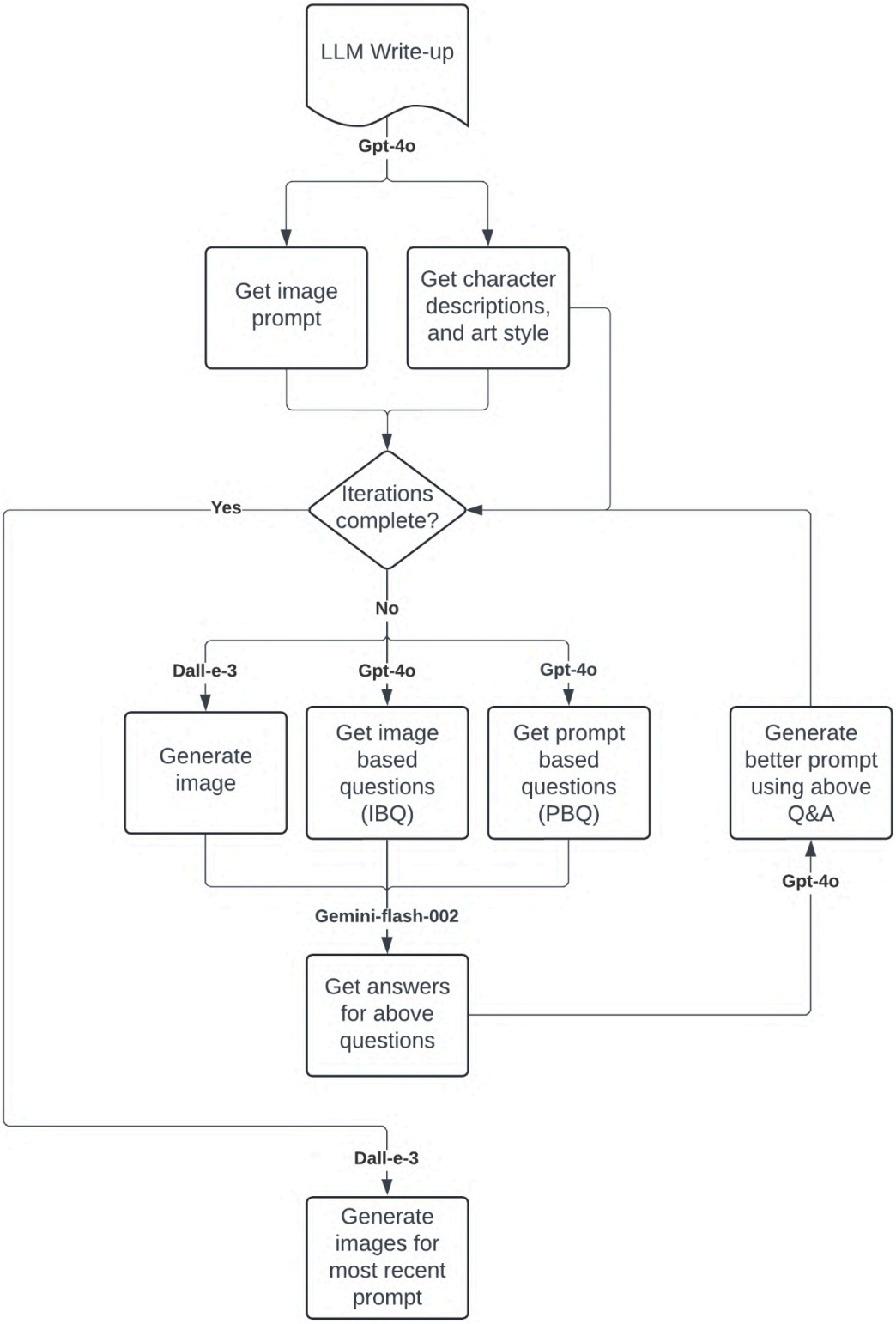
Step 6: Iterative Refinement

Based on the answers from Prompt Verification Questions (PVQs) and Image-Based Questions (IBQs), the original prompt is refined to better align with the desired outcome. Key insights are used to identify and address issues such as:

- Inconsistencies in character representation: Adjusting details like appearance or role if they do not match the image.
- Missing details: Adding specific elements (e.g., setting or actions) that were overlooked.
- Art style misalignment: Clarifying or adjusting the style description if it doesn't match the generated image.

These changes ensure the prompt more accurately reflects the intended characters, setting, and style, leading to better-aligned image generation in subsequent iterations.

Architecture diagram:



Outcome:

Iterative Refinement Results and Model Improvements

1. Initial Attempt:

We used dall-e-3, GPT-4o-mini, and Gemini-1.5-Flash-002, running the iterative cycle twice to refine the prompt. The process involved generating images, having Gemini answer PVQs and IBQs, and adjusting the prompt based on those responses. While the images were decent, the character choices felt somewhat generic and didn't fully capture the essence of the prompt.

2. Enhanced Iteration:

We switched from GPT-4o-mini to GPT-4o, hoping to improve the image quality and character representation. Running the cycle twice again resulted in a noticeable improvement. The images were better aligned with the prompt, and the model handled more nuanced details, leading to more relevant and cohesive character choices.

3. Refined Focus:

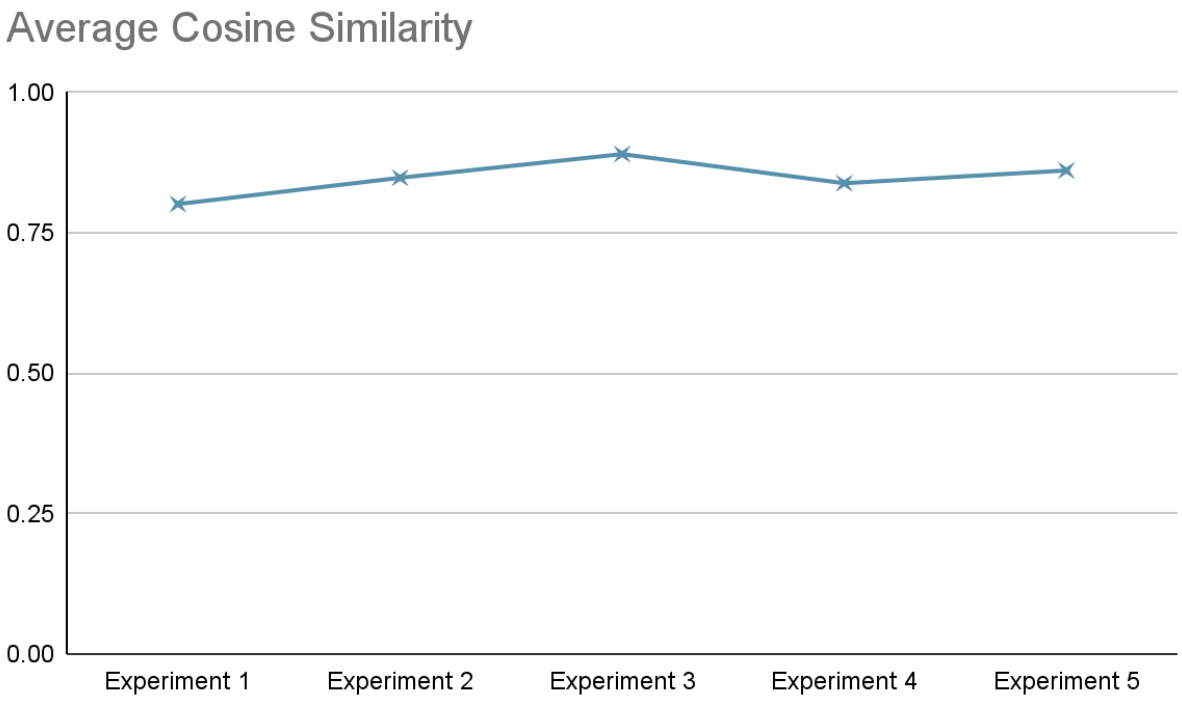
Next, we focused on refining the prompt itself. We shortened it and added character descriptions at the start, which were then appended to each image generation prompt. This modification also ran for two cycles and produced much better results. The images now featured characters that were more relevant and consistent with the scene, and the reduced prompt length allowed the process to focus more on the key elements.

4. Further Streamlining:

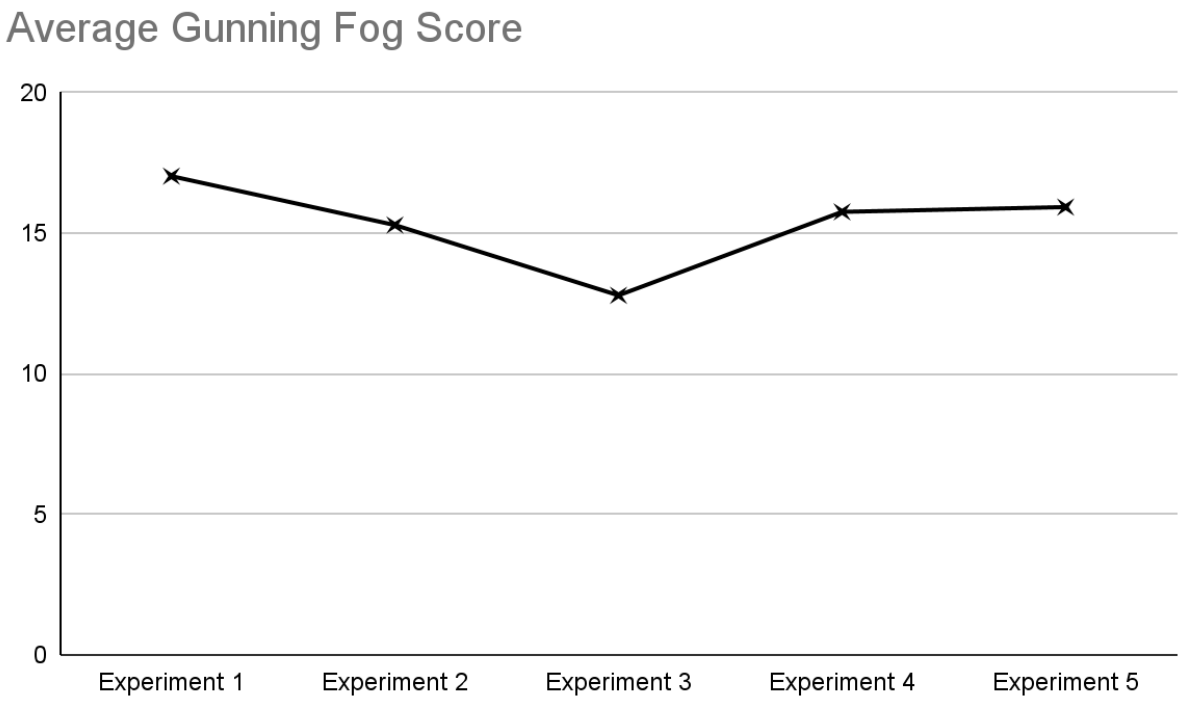
We made further refinements by shortening the prompts even more and asking for a more specific exclusion of background and facial expressions in the character descriptions. We also clarified the art style. Unfortunately, the longer combined prompts caused execution issues, with the model struggling to handle the complexity, which limited the effectiveness of the cycle.

5. Optimal Results:

Finally, we settled on running the cycle four times, aiming for even more streamlined prompts. We emphasized the core aspects of each prompt, clarified the art style, and reinforced the exclusion of background/expression details. This approach yielded the best results yet: the character choices were spot on, and the images became far more relevant to the intended prompt and topic. We also tested GPT-4o-mini (image shown as experiment 4.5) briefly for character descriptions, but its results were less relevant, confirming GPT-4o as the preferred model for the task. The iterative refinement process proved to be highly effective, gradually improving the alignment with the desired outcomes across the cycles.



Above is the variation of the average cosine similarity across the three prompts for each experiment for the Sukanya Samriddhi Account Scheme (SSAS). Below is the variation of the average Gunning Fog score across prompts for each experiment on the same scheme.



The metrics across the experiments reveal a clear progression in prompt refinement, aligning with the iterative approach described. Early experiments, such as Experiment 1, show lower cosine similarity scores (0.8013), reflecting less cohesive prompts and character descriptions. These scores improve significantly in later iterations, with Experiment 3 achieving the highest similarity (0.8903), indicating more consistent semantic alignment. By Experiment 5, the similarity stabilizes at a strong 0.8611, demonstrating the balance between thematic alignment and prompt specificity.

Token lengths and lexical diversity further illustrate this refinement. Early experiments exhibit wide variations in token lengths and higher lexical diversity, which align with the evolving efforts to improve clarity and detail. In contrast, Experiment 5 demonstrates more stable token lengths and reduced diversity, reflecting the streamlined and focused prompts developed during later cycles. This shift aligns with the goal of emphasizing the core aspects of the prompts, as detailed in the refinement process.

Readability metrics support this trend, with Flesch Reading Ease scores declining slightly in later experiments as the prompts became more descriptive and complex. The Gunning Fog scores remain relatively high but consistent, with Experiment 5 balancing clarity and detail effectively. This reflects the success of shortening prompts while focusing on key elements like character descriptions and art styles.

Experiment 5 was ultimately chosen because it represents the optimal point in this iterative refinement process. It balances semantic alignment, clarity, and focus, achieving stable and meaningful results. The streamlined prompts in Experiment 5 allowed for stronger character representation and more relevant images, addressing issues of generic choices seen in earlier cycles. The decision to emphasize core aspects, avoid excessive background detail, and focus on clarity validated Experiment 5 as the most effective strategy for achieving the intended outcomes.

Image Evolution (Last iteration for each experiment)

1. Experiment 3

2. Experiment 4



3. Experiment 4.5 (gpt-4o-mini)



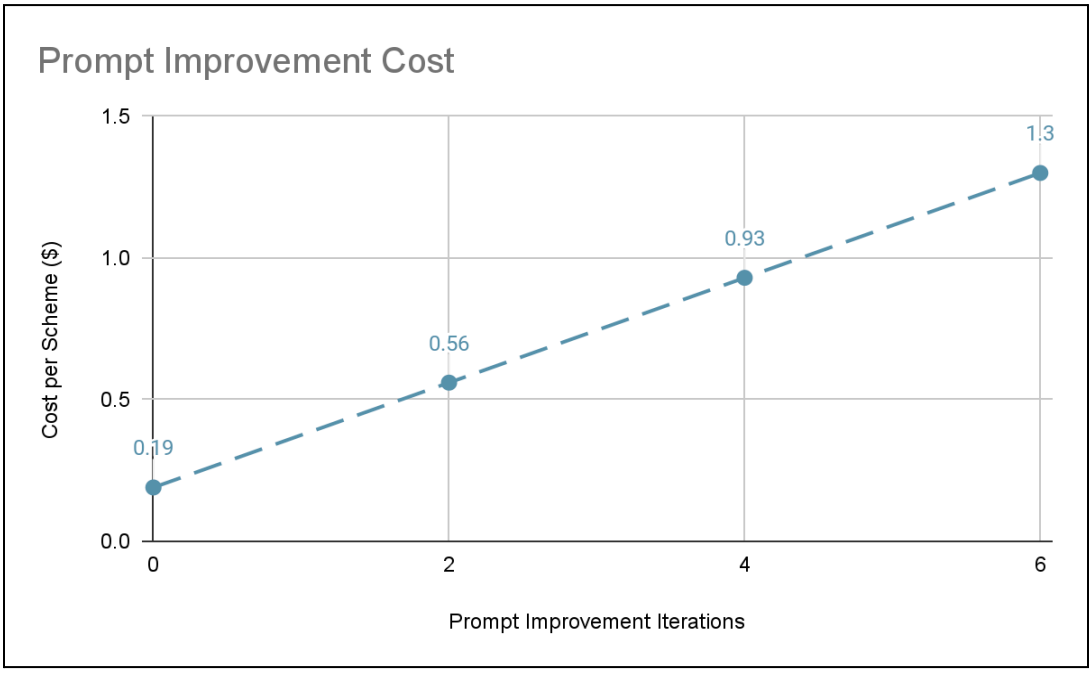
4. Experiment 5



5. Final Output

Cost Breakdown

Each base prompt generation and image generation cost was 19 cents, a slight increase compared to the previous submission where the cost was around 12-13 cents. This rise is due to the inclusion of character description generation and the switch to GPT-4o, which is significantly more expensive than GPT-4o-mini, the model used previously. In addition, refining the prompt once added an extra 18.5 cents to the total cost.
Thus, for a scheme that undergoes two iterations, the total cost of image generation comes to 56 cents per scheme (19 cents for the initial image generation and 18.5 cents for each prompt refinement). When the cycle is extended to four iterations, the total cost per scheme increases to 93 cents.



Conclusion

Our project successfully applied the principles from GenAssist: Making Image Generation Accessible to develop a robust system for generating culturally relevant, coherent, and clear visual content for welfare schemes in India. By addressing challenges specific to this context, such as cultural sensitivity, text clarity in Indian languages, and ensuring consistency in character and art style, we were able to improve upon existing systems in generating visuals that effectively communicate welfare scheme information.

Through an iterative refinement process involving detailed prompt generation, prompt verification, and visual quality assessments, we were able to significantly enhance the quality and relevance of the generated images. By leveraging advanced models like DALL-E-3, GPT-4o, and Gemini-flash-002, we ensured that each iteration brought us closer to the desired outcomes, ultimately achieving a high level of accuracy in terms of visual coherence, entity consistency, and alignment with the original prompt.

The automation pipeline we developed further ensures that this process can be scaled efficiently, making it easier to generate a large volume of customized visuals for various welfare schemes. This will help bridge the communication gap in underserved communities, ensuring that welfare schemes are more accessible and engaging to the target audience.

Despite challenges in text legibility and occasional inconsistencies in art style, our project demonstrates the potential of combining advanced AI models with iterative feedback loops to refine image generation for specific contexts, and at reasonable costs (~₹80 per scheme).