Using In-context Learning to Automate AI Image Generation for a Gamified Text Labelling Task

Fatima Althani, Chris Madge, Massimo Poesio
Queen Mary Univ. Of London, United Kingdom
{f.althani, c.j.madge, m.poesio}@qmul.ac.uk

Abstract

This paper explores a novel automated method to produce AI-generated images for a text-labelling gamified task. By leveraging the in-context learning capabilities of GPT-4, we automate the optimisation of text-to-image prompts to align with the text being labelled in the part-of-speech tagging task. As an initial evaluation, we compare the optimised prompts to the original sentences based on imageability and concreteness scores. Our results revealed that optimised prompts had significantly higher imageability and concreteness scores. Moreover, to evaluate text-to-image outputs, we generate images using Stable Diffusion XL based on the two prompt types, optimised prompts and the original sentences. Using the automated LIAON-Aesthetic predictor model, we assigned aesthetic scores for the generated images. This resulted in the outputs using optimised prompts scoring significantly higher in predicted aesthetics than those using original sentences as prompts. Our preliminary findings suggest that this methodology provides significantly more aesthetic text-to-image outputs than using the original sentence as a prompt. While the initial results are promising, the text labelling task and AI-generated images presented in this paper have yet to undergo human evaluation.

Keywords: Games-with-a-Purpose, Prompt engineering, Generative AI

1. Introduction

Games-with-a-Purpose (GWAPs) for Natural Language Processing face the challenge of engaging players, primarily due to the lack of visuals, unlike their image-labelling counterparts (Lafourcade et al., 2015). Efforts to integrate visuals in GWAPs for NLP have been achieved by developing themes and designing virtual worlds. Nonetheless, while these visuals can be aesthetically appealing, they often fail to support the text being labelled directly. This challenge stems from the resource-intensive nature of creating relevant visuals to accompany each text that requires labelling. To address the lack of contextually relevant visuals in text-labelling GWAPs, we propose a novel approach that uses GPT-4's in-context learning capability to automate AI image generation for text-labelling games. This method included an exploratory approach to selecting a set of in-context tasks to generate an optimised prompt based on our part-of-speech tagging task's word and sentence pairs.

In this preliminary study, we evaluated both the optimised prompts and AI-generated images. To assess the optimised prompts, we measured both the average imageability and concreteness scores, comparing them with the original sentences. We then generated images using Stable Diffusion XL, the latest version of Stable Diffusion (Rombach et al., 2022), using both optimised prompts and the original sentences. Subsequently, the text-to-image outputs were evaluated using the latest version of the LAION-Aesthetics predictor model (Schuhmann et al., 2022) and assigned aesthetic scores for each generated image. After collecting all scores, we conducted a correlation analysis to compare imageability and concreteness with predicted aesthetic scores.

Based on previous work, we hypothesised that:

- **H1** Optimised text-to-image prompts will have higher imageability and concreteness scores compared to the original sentences.
- **H2** Text-to-image outputs of optimised prompts will receive higher predicted aesthetic scores than those generated from original sentences.
- **H3** The higher the imageability and concreteness scores of a prompt, the higher the aesthetic score of the text-to-image output.

The main findings of our preliminary study include:

- The design of a gamified text labelling task that features contextually relevant AI-generated images.
- A description of the iterative process we employed to produce the AI-generated images.
- A quantitative evaluation for both optimised prompts and their text-to-image outputs.

Finally, the main advantage of applying this method is the ability to generate context-relevant images for a text labelling task by utilising LLM's in-context learning abilities. This approach is readily accessible, as it requires designers to iteratively develop a set of instructions for the LLM without

Games and NLP @LREC-COLING-2024, pages 21–31
20-25 May, 2024.

using additional models to reach a desired output. Following this preliminary study, we aim to conduct future studies on how these images impact user engagement in the gamified text labelling task introduced in this paper.

2. Related Work

2.1. Games-with-a-Purpose for NLP

It has been suggested that the inherent nature of the task being text-based is the reason these GWAPs are not as successful (Lafourcade et al., 2015). Visuals are an essential aspect to consider when designing games. For that reason, GWAPs in this domain have found ways to incorporate visuals into their games despite the in-game tasks being mainly text-focused. Previously, several GWAPs have acknowledged and explored using images as a proxy, retrieving them from existing sources to support annotation (Jurgens and Navigli, 2014; Vannella et al., 2014). Moreover, GWAPs for NLP frequently applied various thematic elements to improve the visuals, such as in Phrase Detectives (Chamberlain et al., 2008), WordClicker (Madge et al., 2019), and Wormingo (Kicikoglu et al., 2019). More recently, designers of GWAPs for NLP explored ways of incorporating text-labelling tasks into virtual worlds, such as LingoTowns (Madge et al., 2022), High School Superhero (Bonetti and Tonelli, 2020) and Stroll-with-a-Scroll (Aliady et al., 2022), creating a more visually appealing experience. While these are interesting approaches to enhance the visual appeal of these games, we would like to explore the use of generative AI in a GWAP for NLP. This novel approach will allow us to generate visuals directly corresponding to the text being labelled.

2.2. Co-designing with Generative AI

With the recent rise of generative AI models, many researchers have begun exploring how to use them as tools to support creativity (Liu et al., 2022). Generative AI models can generate text (e.g., GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), LLaMA (Touvron et al., 2023)), images (e.g., Stable Diffusion (Rombach et al., 2022), Midjourney[1], DALL-E (Ramesh et al., 2021), music (e.g., MusicGen (Copet et al., 2024)), video (e.g., VideoGen (Li et al., 2023)). This range of creative capabilities allows generative AI models to assist in design and content creation (Antony and Huang, 2023). Due to the extensive domain knowledge that LLMs possess, they can be suitable for tasks that require knowledge in different fields and domains. This makes using LLMs ideal for designing visuals for GWAPs for NLP, as these games use a wide range of corpora covering various topics.

2.3. Text-to-Image Generation

Text-to-image generation has evolved significantly, starting with Generative Adversarial Networks (GANs) and advancing to Conditional GANs. Within recent years, the development of diffusion models has further increased the popularity of text-to-image models, as evidenced by the widespread use of DALL-E (Ramesh et al., 2021), Midjourney and Stable Diffusion (Rombach et al., 2022). With this increase in popularity, the application of text-to-image models is expanding, particularly in the domain of game design (Deckers et al., 2023). Nevertheless, while this technology offers an accessible and cost-effective way to create game assets, producing quality text-to-image outputs can be challenging. This is due to text-to-image models' output quality greatly depending on the prompt used to generate an image. This has led many researchers (Liu and Chilton, 2022; Oppenlaender, 2023a) to investigate how to craft better prompts for text-to-image generative models.

2.4. Prompt Engineering

Following the popularity of large language models, a novel paradigm of human-AI interaction has emerged, known as prompt engineering (Brown et al., 2020). This practice has evolved into a form of art (Oppenlaender, 2022; Reynolds and McDonell, 2021), where prompt engineers creatively craft a set of instructions in order to achieve a desired output from an LLM. This practice first emerged in the field of NLP, and its usage was extended to text-to-image models. Prompt engineering follows an iterative cycle where prompts are modified and refined after every output until the desired results are reached. This process of manually generating prompts can be laborious. This is especially true when tasked with producing suitable text-to-image prompts that generate images which support the text in a text labelling task. Thus, a more efficient process of prompt engineering is required.

2.5. Prompt Optimisation

Different tools and models have been developed to assist in generating prompts to ease the process of prompt engineering. Prompt optimisation can be a partially manual or a completely automated process. Some methods opting for more of an exploratory approach to prompt optimisation include tools like Promptify (Brade et al., 2023) and Opal (Liu et al., 2022). Both tools use LLMs to guide users into producing improved text-to-image

[1] https://www.midjourney.com/

prompts. Current models that automate prompt optimisation for text-to-image prompts include BeautifulPrompt (Cao et al., 2023), Promptist (Hao et al., 2023). These models apply various automated scoring systems in their prompt optimisation models to improve prompts. These scoring systems include CLIPscore (Hessel et al., 2022), LAION-Aesthetic predictor (Schuhmann et al., 2022), and PickScore (Kirstain et al., 2023). Recently, LLM-score (Lu et al., 2023) was developed to evaluate text-to-image output using LLMs, focusing on the composition of the generated images. Another example of the use of LLMs in prompt optimisation is LLM-grounded diffusion (Lian et al., 2023). This method uses LLMs to represent objects in complex prompts, accurately enhancing the generated image. Another method to optimise prompts using LLMs is through the use of their in-context learning capabilities. This approach has been previously explored to enhance the representation of Arabic culture in generated images using LLMs' domain knowledge (Elsharif et al., 2023). Setting in-context learning instructions is a simple and easily accessible approach to optimising prompts, only requiring the design of a set of tasks for an LLM to follow. This is why we selected this approach to optimise the text-to-image prompts for our task.

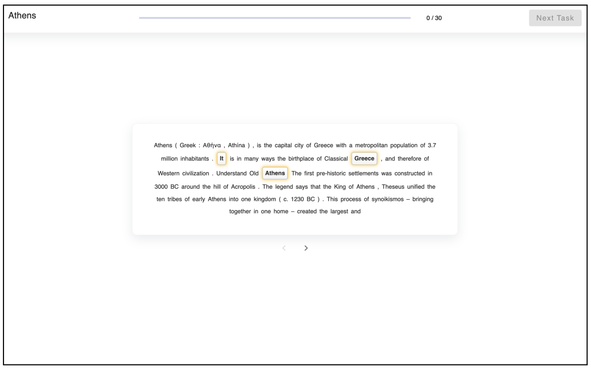
2.6. Evaluating Text-to-Image Prompts and Outputs

The images generated by text-to-image models can be evaluated across various aspects such as text-image alignment (e.g. CLIPscore (Hessel et al., 2022)), aesthetics (e.g. LAION-Aesthetic predictor (Schuhmann et al., 2022)), quality (Salimans et al., 2016) and bias (Bianchi et al., 2023). The LAION-Aesthetic predictor (Schuhmann et al., 2022) was used to evaluate aesthetics in several prompt optimisation models mentioned earlier, such as Promptist (Hao et al., 2023) and BeautifulPrompt (Cao et al., 2023). The same aesthetic measure was used to evaluate and compare a selection of text-to-image models (Lee et al., 2023). Furthermore, text-to-image prompts necessitate the use of visual language (Qiao et al., 2022), making it compelling to evaluate both the imageability and concreteness of a prompt.

3. Design

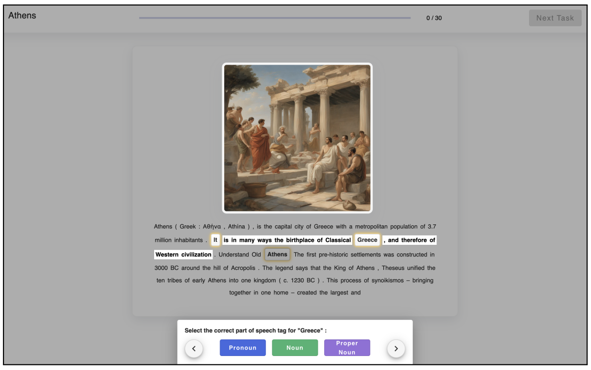
3.1. Interface of the Gamified Text Labelling Task

We developed our gamified text labelling task as an HTML5/Typescript web application using Angular, focusing on labelling nouns, proper nouns, and pronouns (Figure 1). To keep the game simple and maintain user focus on the labelling task, we implemented simple mechanics. Users initiate labelling by clicking on glowing buttons representing words that require labelling. Upon selection, the sentence containing the word is highlighted, and an AI-generated image relevant to the text appears above, accompanied by a bottom sheet displaying part-of-speech tags in different colours: blue for pronouns, green for nouns, and purple for proper nouns (Figure 2). Correctly labelled words change colour to match their part-of-speech tag, while incorrect choices turn the word grey. The interface facilitates seamless navigation to subsequent words, allowing users to label all words efficiently. Additionally, navigation buttons are present to allow users to move between words and progress to the next task. The interface also features a progress bar at the top, displaying the number of words left to label alongside the document's title, ensuring users can easily track their progress.



Athens (Greek : Αθήνα , Athína) , is the capital city of Greece with a metropolitan population of 3.7 million inhabitants . It is in many ways the birthplace of Classical Greece , and therefore of Western civilization . Understand Old Athens The first pre-historic settlements was constructed in 3000 BC around the hill of Acropolis . The legend says that the King of Athens , Theseus unified the ten tribes of early Athens into one kingdom (c. 1230 BC) . This process of synoikismos – bringing together in one home – created the largest and

Figure 1: The interface of the gamified text labelling task.



Athens (Greek : Αθήνα , Athína) , is the capital city of Greece with a metropolitan population of 3.7 million inhabitants . It **is in many ways the birthplace of Classical** Greece **, and therefore of Western civilization** . Understand Old Athens The first pre-historic settlements was constructed in 3000 BC around the hill of Acropolis . The legend says that the King of Athens , Theseus unified the ten tribes of early Athens into one kingdom (c. 1230 BC) . This process of synoikismos – bringing together in one home – created the largest and

Select the correct part of speech tag for "Greece" :

Figure 2: The bottom sheet is displayed after a user selects a word.

3.2. Corpus

We selected six documents from the GUM corpus (Zeldes, 2017), a sizeable open-source multilayer corpus labelled and annotated by experts. The same corpus was previously used to evaluate accuracy in a GWAP for NLP (i.e. WordClicker (Madge et al., 2019)). The selected documents included a mix of fiction and non-fiction texts, offering a diverse sample for analysis. For each document, 30 words were randomly selected, focusing exclusively on pronouns, nouns, and proper nouns. This resulted in a total of 180 words, with an even distribution of 60 words per part-of-speech category.

4. Generation Methodology

This section covers the generation methodology of both the optimised prompts and the AI-generated images.

4.1. Automating AI Image Generation

We followed a two-step process to automate the generation of AI images for our text labelling task. The first step involved optimising prompts based on the original sentences from the corpus. Once complete, the second step was to generate images from the optimised prompts.

4.1.1. Optimising Prompts using GPT-4

To automate the prompt optimisation process, we developed a Python script that iterates over each word and sentence pair in the documents mentioned earlier in Section 3.2. Utilising the GPT-4 model via the OpenAI API, we instructed the model to generate prompts based on a series of in-context learning tasks. The temperature of the model was set to 0 to ensure that the model's output remains deterministic. Before deciding on the final set of in-context learning tasks, we explored multiple sets of instructions through trial and error. This was done by qualitatively evaluating the final outputs of text-to-image that were generated using GPT-4 optimised prompts until the desired output was reached. We initially focused on improving subject coherency by ensuring that the subject was being represented accurately based on the context of the sentence in the document. Following our first set of results, we realised that the setting of that subject was also essential to include in the set of in-context tasks in order to capture the context entirely.

This exploration led us to decide on three tasks for GPT-4 to complete (see Appendix A for the complete set of instructions used). Firstly, the model was tasked with identifying the subject being referred to from the given word based on the context of the sentence in the document. Secondly, we asked the model to describe the setting of the subject identified. Finally, based on the identified subject and setting, GPT-4 generated a text-to-image prompt. The model was instructed to keep the prompt one sentence long, focusing on visual elements while avoiding overly complex language. Using this method, we ended up with a total of 180 optimised prompts. To evaluate the optimised prompt, we had two prompt types:

- Original Sentence: As a control measure, images were generated using unaltered sentences from the documents.
- Optimised Prompt: These prompts were generated using GPT-4 by utilising the in-context learning instructions mentioned above.

4.1.2. Generating Images with Stable Diffusion XL

Each 1080 x 1080 pixels image was generated on either a remote A100 or V100 GPU using the default settings of Stable Diffusion XL (SDXL) 1.0 base model [2] and then refined through the refiner[3]. The refiner uses an img2img approach to improve the image quality. Images were all generated using the same seed 1040 to remain consistent. This resulted in a total of 360 images generated based on two prompt types—optimised prompts and original sentences, with 180 images for each type. It is important to note that for the original sentence prompt type, selecting a word from the same sentence could generate duplicate images.

5. Evaluation Methodology

This section explains the evaluation metrics used to measure the concreteness and imageability scores for the prompt types and the aesthetics score used to evaluate the text-to-image outputs.

5.1. Concreteness

We calculated the average concreteness score of the two prompt types based on the sum of all words' concreteness ratings from the Brysbaert et al. (2014) database divided by the number of words in the sentence to get the average concreteness score. Words not found in the vocabulary were assigned a score of 0.

5.2. Imageability

We calculated the average imageability score of the two prompt types based on the sum of all words' imageability ratings from the MRC database (Coltheart, 1981) divided by the number of words in the sentence to get the average imageability score.

[2] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

[3] https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0

Words not found in the vocabulary were assigned a score of 0.

5.3. Aesthetics

To evaluate the aesthetic appeal of the text-to-image outputs, we utilise the latest version [4] of the LAION-Aesthetic predictor (Schuhmann et al., 2022) which is an automated method for evaluating the aesthetics of AI-generated images. The predictor, trained with human ratings from the Aesthetic Visual Analysis dataset (Murray et al., 2012), predicts the aesthetic scores for images on a scale from 1 to 10.

6. Results and Discussion

In this section, we explore the implications of our findings and discuss how they align with our hypotheses.

- **H1** *GPT-4 optimised text-to-image prompts will have higher imageability and concreteness scores compared to the original sentences*

Our results support our first hypothesis, as imageability and concreteness scores were significantly higher in the optimised prompts. Imageability results from a one-tailed independent sample t-test revealed that GPT-4 optimised prompts ($M = 365.90$, $SD = 36.42$) scored significantly higher in imageability compared to original sentences ($M = 348.65$, $SD = 35.0$), $t(358) = 4.576$, $p < .001$, Cohen's d $= 0.482$ (shown in Figure 3). This demonstrates that GPT-4 optimised prompts are significantly more imageable than original sentences, with a moderate effect size.

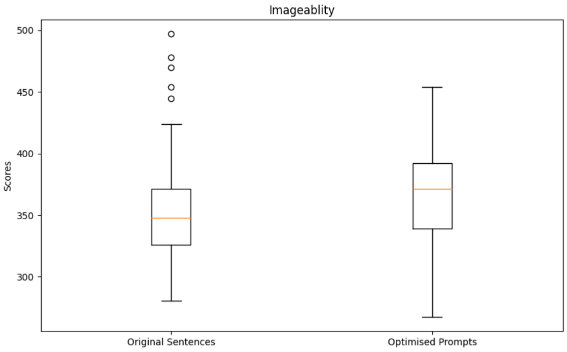


Figure 3: Comparison of imageability scores.

Concreteness results from a one-tailed independent sample t-test showed that GPT-4 optimised prompts ($M = 2.75$, $SD = 0.30$) were perceived to have higher concreteness than original sentences ($M = 2.65$, $SD = 0.31$), $t(358) = 3.129$, $p < .001$, Cohen's d $= 0.330$. This indicates a significant difference in concreteness, favouring GPT-4 optimised prompts over original sentences, with a small to moderate effect size (shown in Figure 4).

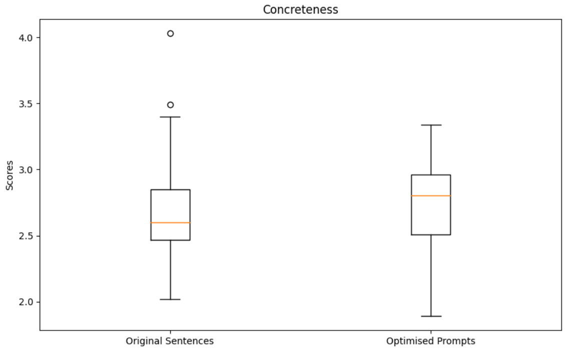


Figure 4: Comparison of concreteness scores.

- **H2** *Text-to-image outputs of the GPT-4 optimised prompts will have a higher predicted aesthetic scores than the outputs using the original sentences as prompts*

Our results support our second hypothesis, as predicted aesthetic scores were significantly higher in the optimised prompts. We evaluated our predicted aesthetics results from a one-tailed independent sample t-test indicated that GPT-4 optimised prompts ($M = 6.29$, $SD = 0.42$) were rated significantly higher in predicted aesthetics than original sentences ($M = 6.07$, $SD = 0.45$), $t(358) = 4.593$, $p < .001$, Cohen's d $= 0.484$. This supports our hypothesis, suggesting a moderate effect size and a significant difference in predicted aesthetics in favour of GPT-4 optimised prompts (shown in Figure 5).

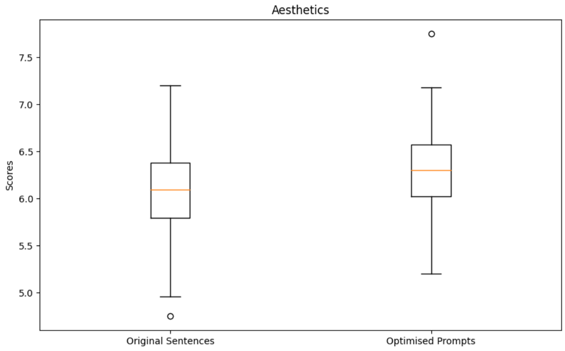


Figure 5: Comparison of predicted aesthetic scores.

- **H3** The higher the imageability and concreteness scores of a prompt, the higher the aesthetic score of the text-to-image output.

[4] https://github.com/christophschuhmann/improved-aesthetic-predictor

Our findings do not support our third hypothesis. Our examination of the relationship between concreteness, imageability and aesthetic scores via Pearson's correlation provided the following insights:

- For concreteness and aesthetics, we found a correlation of -0.127 with a p-value of 0.090, suggesting a weak, negative relationship that was not statistically significant.
- For imageability and aesthetics, we found a correlation of -0.141 with a p-value of 0.058, suggesting a weak, negative relationship that approached but did not reach statistical significance.

These results suggest a nuanced relationship between the imageability and concreteness of prompts and the predicted aesthetics of the generated images, meriting further investigation.

Furthermore, by qualitatively observing the results, we identified apparent differences between the text-to-image outputs generated from optimised prompts versus those from original sentences as prompts (see Appendix B for examples). Optimised prompts, through more descriptive language, consistently resulted in images with vivid colours and detailed compositions. In contrast, images from original sentence prompts often appeared less vibrant and more generic.

The use of GPT-4 to optimise prompts ensures the model more accurately understands the subject or term, effectively improving the ambiguous language found in some original sentences. This led to a noticeable enhancement in the relevance and accuracy of the generated images, addressing issues such as misrepresentation of subjects or settings. These qualitative observations underline that optimised prompts facilitate more accurate and coherent subject representation in text-to-image outputs compared to original sentences as prompts.

7. Limitations and Future Work

Some of the limitations of our study include focusing on only two generative AI models: GPT-4 and Stable Diffusion XL. We did not extend our investigation to other text-to-image models, like DALL-E and Midjourney. Expanding evaluation to other text-to-image models may be necessary, as each model excels in specific areas (Lee et al., 2023). Furthermore, the text-to-image outputs from our study may carry inherent biases, including social and gender biases, as mentioned by Cho et al. (2023). This highlights the need for careful consideration when selecting AI-generated images. Our analysis of text-to-image output was constrained by using a single seed for generating images, limiting our output diversity. Future studies could benefit from generating and evaluating multiple images using different seeds. Tools like PickScore (Kirstain et al., 2023) might be employed to identify the image preferred by users automatically. It is crucial to acknowledge that automated scoring models inherit biases based on their training data. This holds particularly true for subjective tasks such as aesthetics rating. Therefore, biases may be present in the LAION-Aesthetic predictor (Schuhmann et al., 2022), attributable to the subjective nature of aesthetics rating.

Our analysis primarily evaluated optimised prompts against original sentences based on their imageability, concreteness, and predicted aesthetic scores. However, we have yet to compare these AI-generated prompts with user-created ones or fully examine the impact of style modifiers on the AI-generated images, which are known to significantly improve subject coherence (Liu and Chilton, 2022; Oppenlaender, 2023b). Furthermore, to fully grasp the effectiveness of AI-generated images, conducting human evaluations is essential. Extending from this preliminary study, our future work will investigate whether AI-generated images can improve user engagement in a text labelling task.

8. Conclusion

Creating visual content for GWAPs for NLP can be time-consuming and costly, undermining the primary objective of these games. Our paper leverages GPT-4, a large language model, to streamline text-to-image prompt optimisation, introducing an automated approach for generating contextually relevant visual content for text labelling games.

9. Bibliographical References

Wateen Abdullah Aliady, Abdulrahman Aloraini, Christopher Madge, Juntao Yu, Richard Bartle, and Massimo Poesio. 2022. Coreference annotation of an arabic corpus using a virtual world game. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 388–393.

Victor Nikhil Antony and Chien-Ming Huang. 2023. Id.8: Co-creating visual stories with generative ai.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *2023 ACM Conference* *on Fairness, Accountability, and Transparency*, FAccT '23. ACM.

Federico Bonetti and Sara Tonelli. 2020. A 3d role-playing game for abusive language annotation. In *Workshop on Games and Natural Language Processing*, pages 39–43.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. BeautifulPrompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–11, Singapore. Association for Computational Linguistics.

Jon Chamberlain, Massimo Poesio, Udo Kruschwitz, et al. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pages 42–49.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation.

Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. 2023. The infinite index: Information retrieval on generative text-to-image models. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, page 172–186, New York, NY, USA. Association for Computing Machinery.

Wala Elsharif, James She, Preslav Nakov, and Simon Wong. 2023. Enhancing arabic content generation with prompt augmentation using integrated gpt and text-to-image models. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, pages 276–288.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning.

David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464.

Doruk Kicikoglu, Richard Bartle, Jon Chamberlain, and Massimo Poesio. 2019. Wormingo: a'true gamification'approach to anaphoric annotation. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–7.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation.

Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. 2015. *Games with a Purpose (GWAPS)*. John Wiley & Sons.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. 2023. Holistic evaluation of text-to-image models.

Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. 2023. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.

Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–17.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation.

Chris Madge, Jussi Brightmore, Doruk Kicikoglu, Fatima Althani, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2022. Lingotowns: A virtual world for natural language annotation and language learning. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, pages 57–62.

Christopher Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2019. Incremental game mechanics applied to text annotation. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 545–558.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE.

Jonas Oppenlaender. 2022. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, Academic Mindtrek '22, page 192–202, New York, NY, USA. Association for Computing Machinery.

Jonas Oppenlaender. 2023a. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, page 1–14.

Jonas Oppenlaender. 2023b. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, pages 1–14.

Han Qiao, Vivian Liu, and Lydia Chilton. 2022. Initial images: Using image prompts to improve subject representation in multimodal ai generated art. In *Proceedings of the 14th Conference on Creativity and Cognition*, C&C '22, page 15–28, New York, NY, USA. Association for Computing Machinery.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 2014. Validating and extending semantic knowledge bases using video games with a purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1294–1304.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A. Appendix A. The set of in-context learning instructions used

These are the set of in-context learning instructions used for this study. We provided the model with contextual information (document, sentence and word) and a list of tasks to complete.

In-context learning instructions for GPT-4

1. Document: {document}
2. Sentence: {sentence}
3. Word: {word}
4. Task:

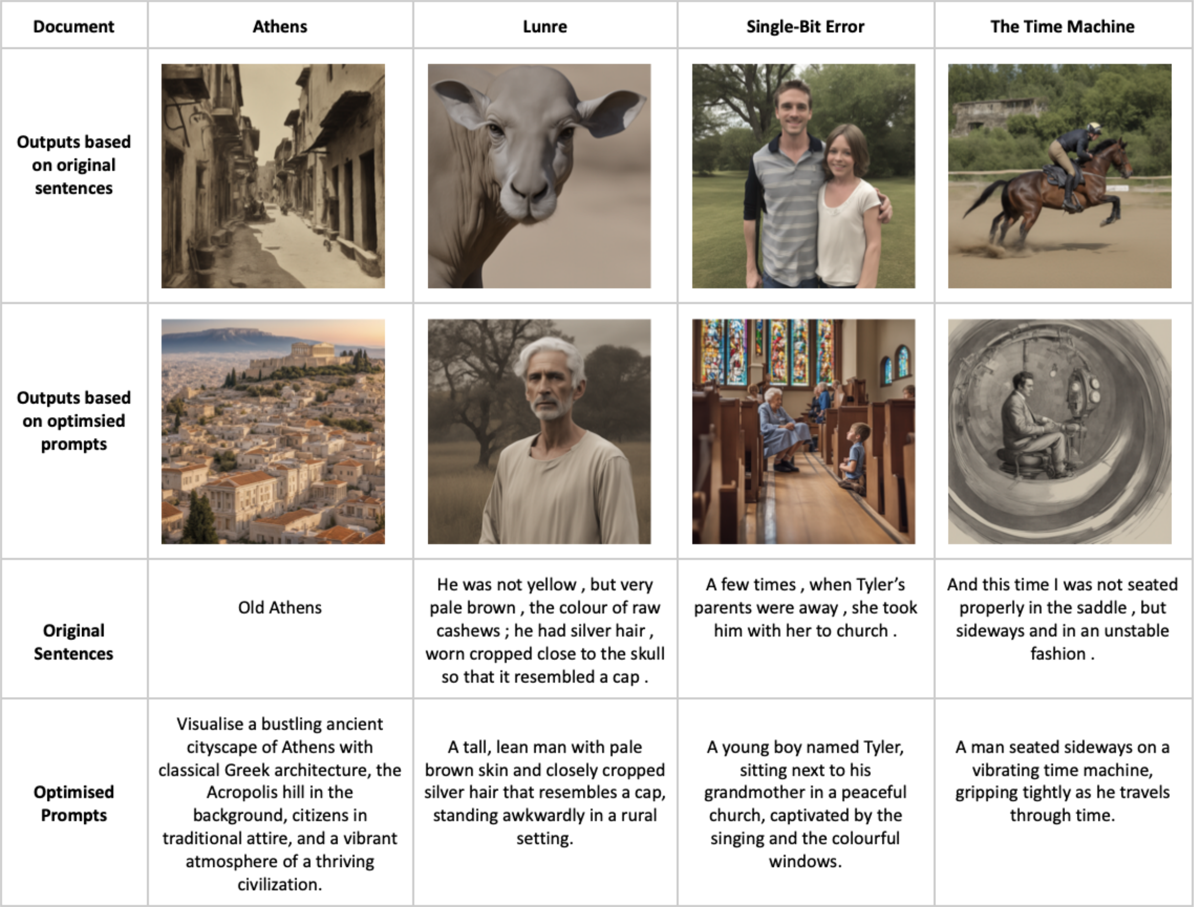
a. Identify the subject being referred to from the given word. The subject can be a character, object, or concept, based on the context of the sentence in the document.

b. Describe the setting of the given subject. This should include the physical environment as well as any relevant mood or temporal aspects.

c. Create a text-to-image prompt that best represents the identified subject and setting. The prompt should be concise yet descriptive, capturing the essence of the sentence or keyword. It should focus on visual elements while avoiding overly complex language. If the sentence is not directly visual, suggest a symbolic or metaphorical representation. This prompt should be one sentence long.

B. Appendix B. Comparison of text-to-image outputs based on prompt types

A qualitative comparison of text-to-image outputs based on prompt types shows how the model fails to present the subjects coherently if the context is not provided. For example, in the Athens document, the optimised prompt, offering more descriptive language, resulted in an image with vivid colours and detailed composition featuring a historical landmark, unlike the original sentence prompt, which produced a dull image of a generic street with old buildings. GPT-4's improvement of ambiguous language is evident in the two following examples. In the Lunre document, "he" mistakenly prompted an image of an animal instead of a man. Additionally, in The Time Machine document, the model incorrectly associated "saddle" with a horse rather than a time machine due to lack of context. Another issue observed with some original sentence prompts was the misrepresentation of subjects' settings. For instance, in the Single-Bit Error document, the text-to-image model did not accurately depict the intended church setting.



Document	Athens	Lunre	Single-Bit Error	The Time Machine
Outputs based on original sentences				
Outputs based on optimsied prompts				
Original Sentences	Old Athens	He was not yellow , but very pale brown , the colour of raw cashews ; he had silver hair , worn cropped close to the skull so that it resembled a cap .	A few times , when Tyler's parents were away , she took him with her to church .	And this time I was not seated properly in the saddle , but sideways and in an unstable fashion .
Optimised Prompts	Visualise a bustling ancient cityscape of Athens with classical Greek architecture, the Acropolis hill in the background, citizens in traditional attire, and a vibrant atmosphere of a thriving civilization.	A tall, lean man with pale brown skin and closely cropped silver hair that resembles a cap, standing awkwardly in a rural setting.	A young boy named Tyler, sitting next to his grandmother in a peaceful church, captivated by the singing and the colourful windows.	A man seated sideways on a vibrating time machine, gripping tightly as he travels through time.