Social Language Strategies of User Prompting for 'Text-to-Image' Deep Learning Models

Keywords: culture and art,artificial intelligence, language, social media

Extended Abstract

Text-to-image synthesis is a field of artificial intelligence that involves using natural language descriptions to generate corresponding images [4]. The quality and diversity of the generated images depends heavily on the input text prompt, making prompt engineering an important aspect of text-to-image generative deep learning and image synthesis [1]. The general goal of prompt engineering is to design effective prompts that guide the generation of high-quality images that accurately represent the intended meaning or content, and is the primary way people interact with this new and quickly developing generative technology.

In a pilot user study to explore prompting strategies, thirty participants selected sculpture-making materials and generated three images using the Stable Diffusion text-to-image generator, each with text prompts of their choice, with the aim of informing and then creating a physical sculpture. The majority of participants (23/30) self-reported that their sculptures were informed by the images they saw, and 28/30 participants reported that they would use text-to-image models again for a creative task.

Using qualitative coding and quantifying how much semantic distance the prompts covered, we identify several prompt engineering strategies. The most minimal amount of conceptual exploration that we observed was a pattern we call the "refiner" style. In these instances, the participant started with a prompt and made minor edits to it. The next pattern observed was the "rephraser" prompting. Here, conceptual subject matter remained the same, but changes to the wording or order of the prompt changed substantially. The "explorer" prompting style we describe consisted of three conceptually unrelated prompts (see figure 1). These described styles form an exploration gradient where "explorers" have the most semantic distance traveled, "rephrasers" have a main idea but still explore around it, and "refiners" focus on exploitation of a single string of words.

We visualize three participant journeys, one from each of the main styles we have described, through semantic space in Figure 2. We use TSNE to reduce dimensionality of the shared embedding space to two dimensions, and draw arrows from points representing a participant's first to second to third prompts. The path between green dots in Figure 2 (representing a single user's prompting journey) shows considerable conceptual exploration using an "explorer" prompting style. The participant journey represented by blue dots shows a less exploratory "rephraser" style, and the user represented by purple dots shows the "refiner" prompting style.

We observed that the distinct prompting styles that emerged differed by the design stage [2] that the participant was in. We found that the amount of conceptual exploration a participant did was lower for participants who said they had a sculpture idea at the visualization stage than those who did not ($t = -2.94$, $p = 0.006$).

To scale this work up 'in the wild', we look at user prompting data from the artmaking social media platform Midjourney. Midjourney users can send their text prompts through the Discord server and receive artistic images back within seconds. The Discord server, as of

September 2022, has over 2.7 million members, and a large majority of the images generated are accessible to the public on their website which receives 3-4 million visitors per month. Approximately 275,000 images are generated by Midjourney on a daily basis (estimate from September 2022) [1]. The Midjourney user prompting data [2] spans a period of 27 days from the summer of 2022, including 145822 user prompts with the correlating timestamps for these prompts and unique user identifiers. The data set also includes tags for whether the prompts were used for 'Upscaling', 'Variations' or whether they were 'initial' prompts. Our proposed method is to use the timestamp data to identify user prompting 'sessions'. A user session can be defined as the time frame within which a user iterates on an idea with a chain of prompt inputs. Identifying such sessions allows us to analyse the chains of prompts sent within a session, and hence identify any emergent patterns of user prompting behaviour. We will analyse the Midjourney data by using CLIP [3] to generate the latent vector for each user prompt, and use the same conceptual distance metric from the pilot study to quantify user's prompting journey. We hypothesise that any user prompting patterns that emerge from the Midjourney data will fall in to a spectrum of identifiable strategies. On one end, the 'explorer' strategy (with large semantic distances travelled within the embedding space), and on the other end the 'refiner' strategy where users iterate on the same string (meaning minimal semantic exploration).

Analysis of prompting data for text-to-image generative deep learning models exposes patterns in user behavior during interactions with this new technology. We have developed a way to characterize this behavior with this novel tool, at scale, by giving a systematic way to characterize prompting behavior- defining language and methods around emerging human behaviors that we now have new forms of trace data for.

References

[1] Vivian Liu and Lydia B Chilton. "Design Guidelines for Prompt Engineering Text-to-Image Generative Models". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501825. URL: https://doi.org/10.1145/3491102.3501825.

[2] Philippa Mothersill and V. Michael Bove. "Humans, Machines and the Design Process. Exploring the Role of Computation in the Early Phases of Creation". In: *The Design Journal* 20.sup1 (2017), S3899–S3913. DOI: 10.1080/14606925.2017.1352892. eprint: https://doi.org/10.1080/14606925.2017.1352892. URL: https://doi.org/10.1080/14606925.2017.1352892.

[3] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *CoRR* abs/2103.00020 (2021). URL: https://arxiv.org/abs/2103.00020.

[4] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[1] https://tokenizedhq.com/midjourney-statistics/

[2] https://www.kaggle.com/datasets/succinctlyai/midjourney-texttoimage

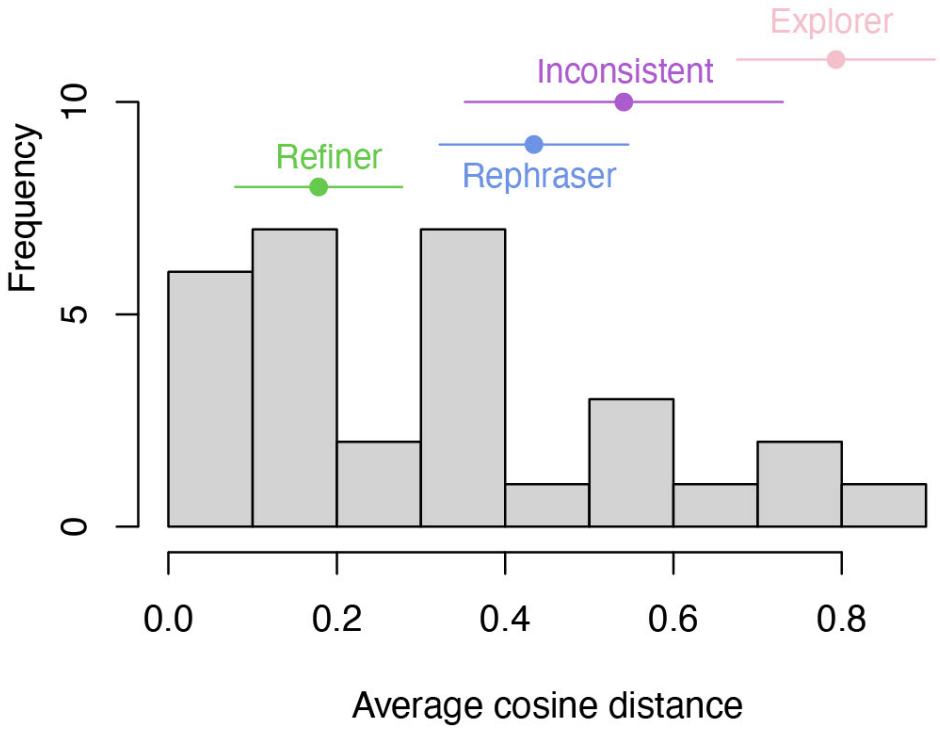

Figure 1: Histogram of participants' conceptual exploration across their prompting session, measured by average co-sine distance. Shown above histogram: mean cosine distance within a prompting style and 95% confidence intervals. The prompting styles appear to occur at different average cosine distances.

.

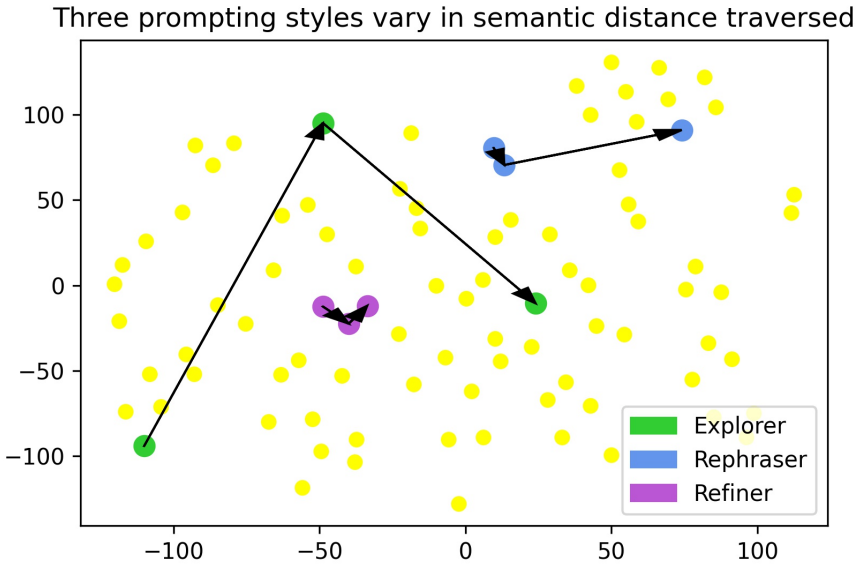

Figure 2: The prompt embedding space (t-SNE) for all users, colour coded for the three prompting strategies - one user per strategy, three prompts per user.

.