Utilizing Prompt Engineering to Enhance Text-to-Text and Image-to-Text Models

Artashes Artashesian, Annette Roberta Petrosyan, Michael Xiong, Cyrus Buffington, Ghirish Thaenraj

Abstract

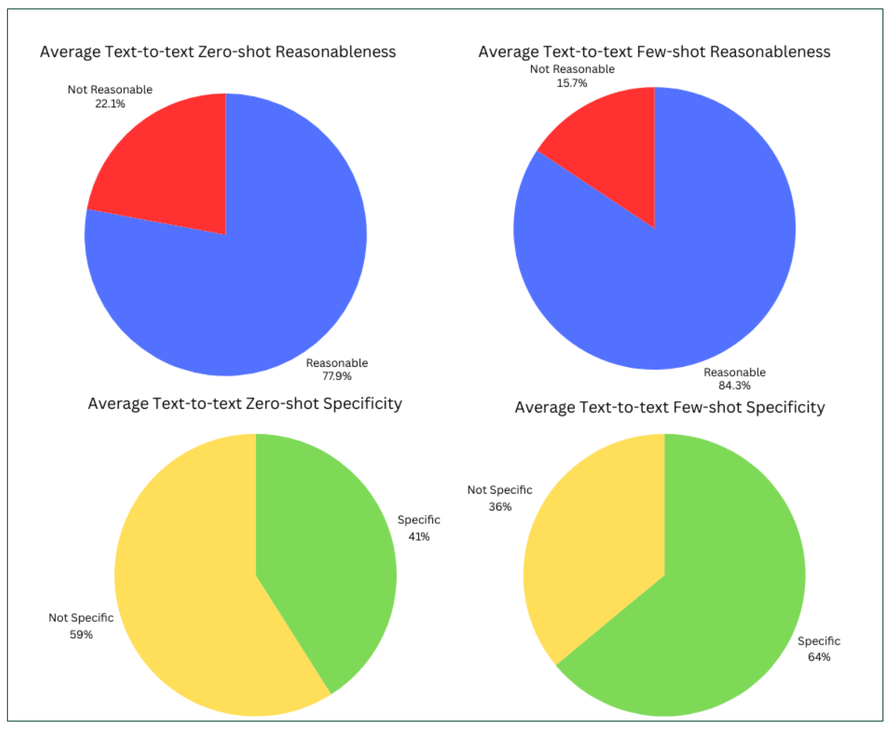
As artificial intelligence (AI), specifically large language models (LLMs), become increasingly prominent in the world of consumer products, technical and non-technical users are looking for ways to get the most out of AI without diving into complex machine-learning concepts. An area of study that explores how to get higher-quality AI responses is prompt engineering. By learning how to speak the language of AI more fluently, users can hone the full potential of AI technologies for both personal and professional applications.

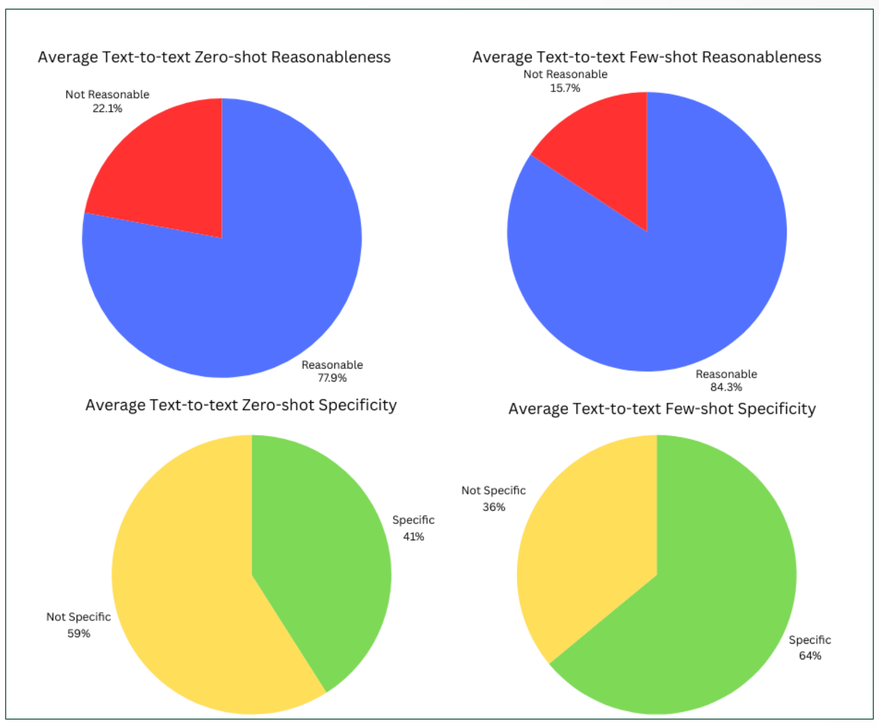
Solution

In this project, we evaluate the effectiveness of prompt engineering methodologies on text-to-text and image-to-text models by comparing the efficacy of zero-shot and few-shot prompting techniques. The quality of LLM output relies heavily on the input users provide, and since users tend to provide short, unspecific prompts, we aim to enhance their experience.

About Data

Users labelled each output as specific or non-specific, reasonable or non-reasonable. We chose “reasonableness” and “specificity” as labels because they allowed us to see if the result was generally factual, aligned with the prompt’s intent, and provided enough detail to be informative without being overly vague or narrow.





Methods

For testing Text-to-Text zero-shot versus few-shot, we used Flan-T5 large, a large language model developed by Google frequently used for Natural Language Processing, and Blip-2, a multimodal transformer that incorporates vision, making it appropriate for Image-to-Text.

We manually adjusted the prompts and key hyperparameters, such as the number of beams, temperature, top-p, and maximum token length, to achieve better results during this process.

During testing for text-to-text, we found that instructing along with the prompt gave the best results for zero-shot and few-shot settings while using Flan-T5 maximizing its abilities.

Semantic Similarity	Prompt 1: Climate Change	Prompt 2: Photosynthesis	Prompt 3: Pythagorean	Prompt 4: Database Indexing	Prompt 5: Cybersecurity	Average Semantic Similarity
Zero Shot vs Few Shot Semantic Similarity	0.6248	0.5135	0.2178	0.4694	0.5222	0.430725
Few Shot vs Prompt Semantic Similarity	0.6170	0.6586	0.2808	0.5884	0.6481	0.543975

Results

Across both experiments regarding the text-to-text and image-to-text tasks, we observed the few-shot prompt consistently yielded better outputs with higher reasonableness and specificity. Our text-to-text few-shot settings performed appropriately, with semantic checks proving the few-shot example prompts are having an influence, but not repeating examples or by being repetitive. Overall, Image-to-Text feedback was the best, showing that few-shot settings performed better on Image-to-Text than for Text-to-Text transformers.

Image-to-Text Results

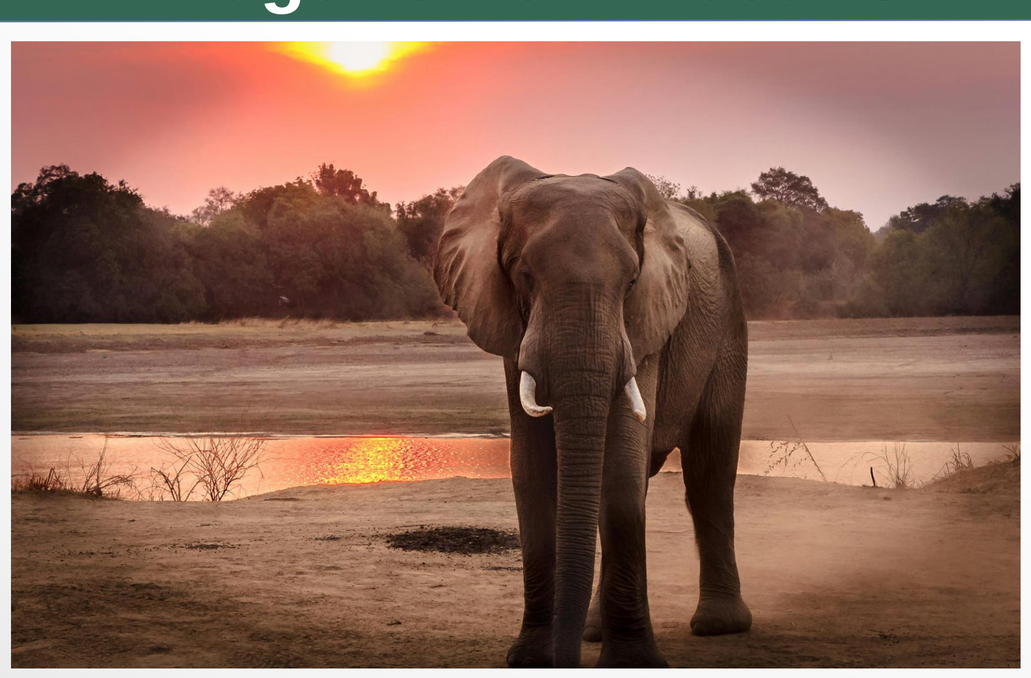


Image-to-Text Results

Question: What is shown in the image?

Zero-Shot: An elephant

**Providing the model descriptive examples such as:

“"There is ... and also ... ",
"An orange basketball rolling on the court and also a player in a red jersey running toward it.",
"There is a steaming cup of coffee on the table and also a plate of croissants beside it.”,
*... etc***

Question: What is shown in the image?

Few-Shot: The sun is setting behind the elephant.

Future Directions

For future directions, we can train a LoRA-adapted version of Flan-T5, where the zero-shot output serves as a prompt for few-shot learning, improving the model’s effectiveness as a prompt generator. For image-to-text tasks, we could transition to the LLaVA multimodal model, which can automatically generate prompts based on image content. This would result in less hard coding of human-written prompts.