# Final Project: AI Text-to-Image Generation

Grace Eunji Kim , Madi Zhaksylyk , Jyoti Girdhari Khanchandani

April 25 2024

Git Repository

Presentation Slides

## Project Overview

Our project aims to generate images from textual descriptions using DALL-E, an advanced text-to-image model. We started by using a Large Language Model (LLM) to generate text descriptions based on product descriptions and reviews. Then, we used DALL-E to convert these textual descriptions into corresponding images. This approach enhances e-commerce platforms' visual content generation capabilities by automating the process of creating images from raw text.

## Data Collection

For data collection,three products were selected from three distinct categories: Travel & Luggage, Kitchen Appliances, and Wall Art.

The first product is the **Samsonite Luggage**, which has core features such as durability, wheels, and lock features, which are essential for frequent travelers. With a customer feedback rating of 4.5 out of 5, based on 5,705 reviews and over 800 purchases in the past month, this product provides a wealth of information on user experiences and product satisfaction. Additionally, the emphasis on simple design and specific features tailored for travel, makes this product particularly interesting for analysis.

The second product is the **KitchenAid Espresso Machine** which includes features such as commercial-grade quality, smart dosing technology, and ease of cleaning. This product also has a high customer feedback rating of 4.7 out of 5, indicating a high level of satisfaction among users. This product was chosen because its descriptions and reviews focused highly on functionality over visuality, making it ideal for analyzing the nuances of product performance and usability.

The third product is the **"Love is the Answer" canvas print** that includes features such as vibrant colors, high-quality printing and material, and ease of hanging. With a customer feedback rating of 4.7 out of 5, based on 831 ratings and 123 reviews, this product offers insights into how consumers perceive and interact with visual art products. The rationale behind choosing this product was to test the ability of DALL-E to capture the imagery and text embedded in the product.

## Technical Overview

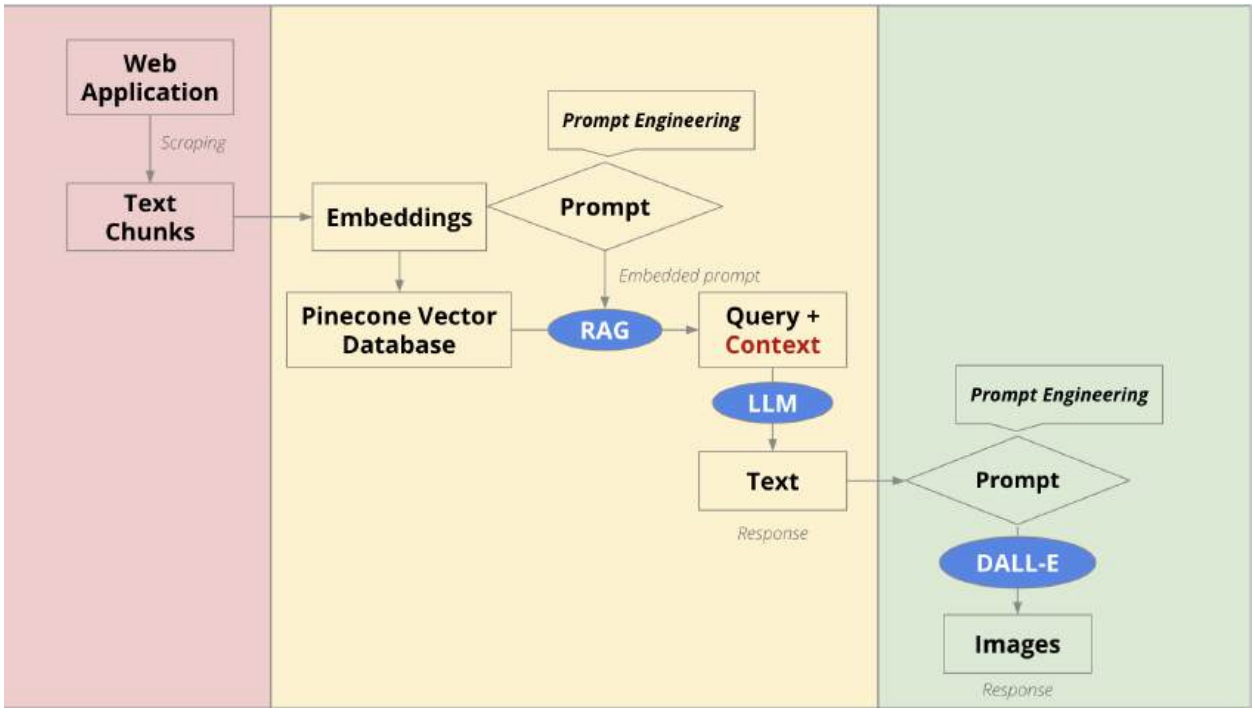


Figure 1. Technical Overview

The data collection process began with the scraping of descriptions and reviews from a web application. These text segments were then embedded using an embedding model and uploaded to the Pinecone vector database. Subsequently, queries were created and embedded to retrieve relevant text chunks from the vector database based on similarity metrics. Using the RAG-generated prompt (query+context), GPT provided responses to questions regarding product features and customer sentiments. The resulting responses were stored and utilized to create the prompt for the DALL-E model for image generation, which also involved prompt engineering. The final output of this process is a generated image.

**Text Analysis using LLM + Image Generation using DALL-E**

**Product 1: Samsonite Omni PC Hardside Expandable Luggage**

Queries and Responses used for LLM

```
["Describe in detail what this product is, based on the product description.",
 "What are some features of this product, especially focusing on visual features?",
 "What is the main purpose of this product?",
 "What is most loved about this product?",
 "What is most hated about this product?"
```

Context chunks received:
*"Great product and love the look. Very good product I bought this luggage for an upcoming 2-week trip. I love the color, but I also did some research before making my purchase. Samsonite makes quality luggage and I read that a **hard-sided luggage** is more **durable** than what I usually use. I hope to have this luggage last me a very long time and I will do it in style. The **radiant pink color** will stand out, make it easy to spot, and it is **pretty**. The luggage is very roomy, has the **expandable feature, locks, and wheel.** I love it! Nothing Haven't used yet, but I'm happy so far with purchase. Like darker color without being black. Good size with expandable sides. I haven't used this yet but looking it over I'm very happy with my purchase. Large suitcase is huge. **Colorful and durable.** I travel frequently. These are great in the 3 different sizes. **Plenty of room.** Wheels and zipper have been durable. Excellent purchase. **Durable**. Love it. Have not had this for long to give a proper review, but so far happy with purchased. Will update if needed. Traveling Perfect I like everything about this except they neglected to place a handle on the wheel end to help with placing them in a car or on a bed to pack and unpack. Someone was sleeping during the design phase. My mother loved it! Nice color as well This is the best carry-on suitcase, really fits so much, very surprised how nice it is. Rugged. Easy to menuver. "*

**Prompt:** Describe in detail what the product is, based on the product description.
*"Based on the provided information, the product in question is a Samsonite hard-sided luggage in a radiant pink color. It is described as durable, roomy, and stylish, with features like expandable sides, locks, and wheels. Customers have expressed satisfaction with the purchase, noting the quality of the product, its size, durability, and vibrant color. Some users mentioned that the suitcase is easy to maneuver, fits a lot of items, and has stood the test of travel. However, a suggestion was made for the addition of a handle on the wheel end for easier handling. Overall, the feedback is positive, highlighting the product's functionality, durability, and aesthetic appeal."*

The response from the LLM successfully summarizes text, extracts features, and performs sentiment analysis. The use of the Retrieval-Augmented Generation (RAG) model was

particularly beneficial, as the retrieved context chunks were highly relevant to the queries. This ensured that the generated responses were not only accurate but also contextually meaningful. Additionally, prompt engineering was employed, requiring specific examples of visual features related to the product, such as color, size, dimensions, material, shape, pattern, and logo. This approach resulted in responses that provided detailed information on these features.

Prompt Engineering and Image Generation with Dall-E

Figure 2. Original Product Image

1. **[Baseline]:** "Generate an image of a product using the following information: [product features description] + [product sentiment]."

Figure 3. Dall-E Generated Images

The generated images are consistent across multiple iterations. All of them include key features such as four wheels, a lock on one side, and a handle at the top. It also performed decently in replicating the material and the color.

2. [Baseline] + “Create an image that illustrates a human being using the product in a real-world setting, showcasing the usage of the product.”

Figure 4. Dall-E Generated Images

This prompt requested an image featuring a person using the product, but most of the generated images appear visibly AI-generated, lacking realism. For instance, there are issues with scale, with the human figure appearing either too large or too small. Additionally, while the image seems to depict an airport background to reflect a "real-world setting," the colors are overly vibrant and animated, detracting from the realistic appearance.

3. [Baseline] + “Photograph, sharp focus, cinematic, sigma 85mm f/1.4 and ISO 100”

Figure 5. Dall-E Generated Images

To enhance realism, this prompt employs a strategy focusing on photorealism. Specific lens specifications were added as modifiers to achieve this effect. The aim is to create an image that closely resembles reality and meets viewers' expectations. This strategy led to images that closely resemble reality, with more nuanced lighting and toned-down colors. However, there was a hiccup - for the image on the right, a camera was humorously added to the corner of the image.

**Product 2: Semi Automatic Espresso Machine with Burr Grinder**

Queries and Responses used for LLM

**Prompt 1:** Describe in detail what the product is, based on product description.
*Based on the reviews provided, the **espresso machine** being discussed seems to be a **KitchenAid** product that has received positive feedback from users **who received it for free from Influenster for review.** Users mentioned that the machine produces coffee shop-like results, has **amazing grinding performance, easy cleaning, and provides a recipe book and clear instructions for use.** It is noted as not being a **simple machine**, but worth the investment for coffee lovers looking to enhance their coffee experience. Overall, users found the machine easy to use, efficient in grinding beans, and a great addition to their daily coffee routine.*

The model correctly identifies the discussed product and brand, along with the positive sentiment attached to it. However, it incorrectly generalizes that feedback is only from users who received the product for free. This generalization, based on some reviews, overlooks the segregation and independence of each review, indicating a lack of understanding in this regard.

**Prompt 2:** What are some of the features of the product, especially focusing on the visual features?
*Based on the reviews provided, the **most loved/liked quality of the espresso machine** seems to be its incredible **customization options and the ability to create a personalized coffee experience tailored to individual preferences**. Users appreciate the machine's **features** that allow them to adjust settings such as **grind texture, volume of grounds, water temperature, and crema**. Despite the **learning curve** involved in understanding and adjusting these settings, users find the process rewarding and enjoy the trial and error involved in customizing their coffee to achieve the **perfect flavor and intensity.** Additionally, the **quality of the espresso** produced by the machine is highly praised, with users comparing it to coffee shop-like results. Furthermore, the durability, build quality, and additional features such as a milk frother/steam wand, underlighting for **clear visibility, and a large water reservoir with a filter contribute to the overall positive experience with the espresso machine.** Overall, the ability to create a high-quality, customizable coffee experience seems to be the most loved quality of the espresso machine.*

Even after specifying “visual features” in the prompt, the response primarily focuses on the functionality as that is what is being mostly described in the description and reviews.

**Prompt 3:** What is the main purpose of this product?

*The purpose of the espresso machine highlighted in the reviews is to provide high-quality espresso and coffee drinks with ease of use, consistent flavor, and a sleek design. Users appreciate the machine for its durability, easy maintenance, ability to grind beans, adjustable settings for grind coarseness and brewing temperature, as well as the convenience of producing single or double shots of espresso at the push of a button. Additionally, users find the machine to be a kitchen staple for coffee enthusiasts, offering the functionality to create a variety of espresso-based drinks quickly and efficiently.*

Since most of the document extracted only discusses this, all prompts and responses lead to the response that seems apt for this prompt. Hence, this response acts as a summarization of the document as well.

**Prompt 4:** What is most loved about this product?
*The product in question is an espresso machine that is highly praised for its ability to make espressos and lattes efficiently. Users love the convenience of being able to grind coffee beans and make espresso shots all in one machine. It is also noted for its ease of maintenance, ease of use, and the beautiful design. Additionally, the machine features Smart Dosing Technology for ideal espresso volume and consistency, a Flat-Base Easy-Tamp Portafilter for stable tamping and optimal flavor, and a Multi-Angle Steam Wand for frothing milk.*

The response mostly consists of positive reviews capturing all the qualities and features of the machine. Beautiful design is mentioned but it is not described.

**Prompt 5:** What is most hated about this product?
*Based on the reviews provided, it seems like the coffee grinder and espresso machine mentioned are* ***praised for their ease of use, high quality, and ability to produce great-tasting coffee.*** *Customers highlight features such as easy setup, detailed instructions, durability, and the ability to customize coffee to personal preferences. The machines are also* ***commended for their modern design and space-saving qualities****. Overall, these products are seen as game changers for coffee lovers and are highly recommended for those seeking a coffee shop experience at home.*

In truth, there are no negative reviews for this product. The expected response from the Large Language Model (LLM) would be to acknowledge this by stating that there are no negative reviews or discussions about the product. However, the LLM continues to focus on the positive aspects, indicating a potential lack of understanding in LLMs.

**Prompt 6:** Provide more details about the visual features.
*Based on the reviews provided, the KitchenAid Espresso Maker seems to be a* ***professional-grade machine with a learning curve*** *but incredible results. It is praised for its customization options,*

*solid* ***build quality, fast operation, easy cleaning, and impressive features such as the milk frother/steam wand.*** *However, some users mentioned that there was a* ***learning curve in understanding the settings and techniques to achieve the ideal cup of coffee. Better visual instructions and clarity on grind size and temperature settings were suggested for improvement.*** *Overall, users are impressed with the machine and anticipate using it for years, considering it a staple in their kitchen.*

Since there are no visual features mentioned in the text, the response does not include any description of the visual features.

**Prompt 7:** What is the product body made of?
*The material of the body of the espresso machine is stainless steel, as mentioned in the review by foxgiirrl.'*

As seen by the response, only one review discusses the material and texture of the product.

The queries were designed in a specific way to prompt the LLM to extract physical characteristics and specifications of the product. However, since most descriptions and reviews focus on the product's functionality rather than its physical attributes, no physical attributes were extracted. Despite this, the LLM accurately captures the functionalities of the product that are praised. The next step is to analyze if DALL-E can generate an image based on these functionalities.

Prompt Engineering and Image Generation with Dall-E

Figure 6. Original Image

1. [Baseline]: Generate an image with the responses from the LLM.

Figure 7. Dall-E Generated Image

The images primarily focus on the functionality of the product, highlighting features like coffee bean storage capacity and water filtering. However, in trying to capture all the minute details, the model overlooks some basic aspects of the real world. For example, it depicts a cup hanging in the air and attaches four boxes of coffee beans to the machine. The model also attempts to label the components but produces gibberish labels. Additionally, it provides two variations: digital and analog display.

2. [Baseline] + “Include more visual features.”

Figure 8. Dall-E Generated Image

The phrase "professional-grade machine" has been interpreted as a professional using the machine in the image. Similarly, "learning curve" with the machine has been translated into a "manual." The resulting image focuses more on the professional and the concept of a learning curve, rather than showcasing the product itself.

3. [Baseline] + “Include material of the body.”

Figure 9. Dall-E Generated Image

The LLM response indicated that the material is "stainless steel," leading to everything in the room being depicted as stainless steel in the image. When new details are introduced, the previous ones are lost. Although the image briefly improves with additional details, it eventually deteriorates, missing the previous details entirely.

4. [Baseline] + “based on product specifications.”

Figure 10. Dall-E Generated Image

The prompts' focus on exact specifications causes the model to grow weary of capturing details repeatedly. As a result, the image includes the equipment laid out as specified in the prompt.

5. **[Baseline]** + "based on visual features *(manually added features).*"

Figure 11. Dall-E Generated Image

In this prompt, visual features were manually described because the LLM responses did not retrieve any. For this product, this approach works best as it captures all of the

machine's physical or visual attributes. The generated images now closely resemble the original image. Additionally, the model assumes digital calibration because it was not specified otherwise.

6. Generate an image of a semi-automatic espresso machine.

Figure 12. Dall-E Generated Image

This image of an espresso machine was created by the model without any additional input or context. It is not specific to the product being analyzed. This image can be used as a point of comparison.

**Product 3: “Love is the Answer” Canvas Print**

Queries and Responses used for LLM

The queries used to pass to LLM captured five different aspects of the product:

1. **Product description and purpose**: “Based on the product description, describe what the product is and its main purpose.”
2. **Potential customer and emotions**: “Describe the potential customer of this product and the emotions associated with it.”
3. **Physical and visual characteristics of the product**: “Describe how does the product look like, focusing on physical and visual features such as size, material, color, odor, taste, weight.”
4. **Specific details or elements of the product**: “Describe any specific details or elements of the product that make it different from similar products, including any phrases, symbols, or messages”
5. **Customer sentiment**: “What is most loved and hated about this product?”

**Prompt 1: Product description and purpose**
*“Based on the feedback, the product description includes phrases such as "Great product," "Amazing quality," "Great canvas," "Exactly as shown," "Love it," "Real nice," "As described," "Perfect," "Great value for the money," and "Beautiful." Customers appreciate the quality, packaging, accuracy of the product compared to the description, and express a desire to buy again or gift it. It seems to be a well-received and highly recommended product.”*

The response we received, which the LLM considered most similar to the query about the product description and purpose, was not as expected. It missed describing what the product is and its primary use. Instead, the response focused on customer reviews, including their judgment of the product's appearance and quality.

**Prompt 2: Potential customer and emotions**
*“Based on the reviews, the product being described is a colorful and vibrant canvas print featuring Albert Einstein in a fun-loving manner. Customers have expressed **great satisfaction** with the quality, size, and brightness of the colors in the artwork. The print has received high praise for its ability to **bring joy and happiness** to those who view it. The overall consensus is that the product is worth the purchase, with many customers mentioning their desire to buy it again or recommend it to others.”*

Again, the LLM response did not directly relate to the query input, especially regarding the customers of the product. However, it did recognize that the canvas print brings joy

and happiness to customers. Interestingly, this response effectively answers the question asked in the first query and specifies that it is *"a colorful and vibrant canvas print featuring Albert Einstein in a fun-loving manner."*

**Prompt 3: Physical and visual characteristics of the product**
*"The product description highlights that the product is of **amazing quality with vibrant colors** and great appreciation from the company. Customers have found it **well-packaged, lightweight, and easy to hang**. The artwork is as described and **looks great in various settings**. It is a **colorful piece** that makes the customers smile and brings joy every time they see it. The **colors are bright, and the quality is fantastic**, making it a beloved piece of art. Despite being described as **lightweight, it is also of good quality**. Customers feel that it is exactly as shown and are satisfied with their purchase, stating they would buy it again or as a gift."*

The LLM successfully retrieved physical and visual characteristics of the product from customer reviews. The response includes various physical dimensions of the product, including quality, weight, colors, and functionality.

**Prompt 4: Specific details or elements of the product**
***"**Based on the customer feedback, the product is described as a great quality item that is appreciated by customers. It adds a **lot of color to the living room**, is exactly as shown in the picture, and **fits perfectly in the office** without looking cheesy or cheap. While the colors may not be extremely vibrant, the piece still **stands out in a room and looks incredible under black lights**. Overall, it is a gorgeous piece that meets expectations and is loved by customers."*

The response did not specify the specific elements of the product that were expected, particularly the sign in Einstein's hands, which was considered the most important part. Instead, the model provided text on how the product fits in the living room or office and its appearance in various surroundings.

**Customer sentiment**
***"**The product is described as being of **amazing** quality, with bright colors and great appreciation from the company, making customers **want to come back for more**. It is **well worth the money, exactly as shown in the pictures**, and the largest photo found of the image. Customers **love the product**, stating that it is their **favorite piece of art** and that it **looks nice** in their living room. Overall, the product is **vibrant and beautiful**, even though some users mention that **it feels a bit cheap in terms of material."***

The LLM response to this prompt was quite accurate, as it meticulously described (mostly) positive and (one) negative customer feedback.

Overall, the queries passed to the LLM were designed to enable effective summarization of the description, extraction of physical features and specific elements, and analysis of customer sentiment. For this product, the model performed relatively well in extracting physical characteristics and customers' feelings, but it struggled to accurately describe the product, its use, audience, and specific elements. This limitation may be attributed to the scarcity or absence of such information in the scraped reviews.

Prompt Engineering and Image Generation with Dall-E


Figure 13. Original Product Image

Having combined LLM-generated responses into context information, we moved to generate images using DALL-E. Based on that contextual information, we crafted prompts to guide the model.

1. **[Baseline]**: “Generate an image of a product using the following information.”


Figure 14. Dall-E Generated Image

The image generated using the basic prompt features a fun-like human face similar to Einstein's, capturing the colorfulness of the product. Additionally, the image includes text indicating the overall sentiment of the reviewers ("Great Product").

2. **[Baseline]** + “Make sure it captures specific elements of the product.”

Figure 15. Dall-E Generated Image

The second prompt instructed DALL-E to focus on specific elements. Since the contextual information highlighted how the product fits into the living room or office and looks incredible under black lights, DALL-E generated a painting featuring Einstein and colorful elements mounted on the wall of a dark room.

3. **[Baseline]** + “Focus on emotions associated with the product.”

Figure 16. Dall-E Generated Image

In another prompt, we directed the model to focus on the emotions associated with the product. However, DALL-E placed excessive emphasis on emotions and created an image featuring elements typically associated with joy and happiness, such as a rainbow, playful animals, a fruitful tree, nature, and bright colors, while completely omitting Einstein.

4. **[Baseline]** + “Make sure to include the phrase that is an important element of the product”

Figure 17. Dall-E Generated Image

To confirm DALL-E's capability, we instructed it to include an important phrase related to the product. As expected, since no such phrase was provided, the model generated a cheerful animated character (emoji) and a colorful, positive aura. It continued to miss

including Einstein in the image. This suggests that without specific contextual data, DALL-E relies on the most prominent features of the product, such as vibrant colors, joy, and positivity.

5. **[Baseline]** + “Include Einstein and his sign with 'Love is the answer' message”

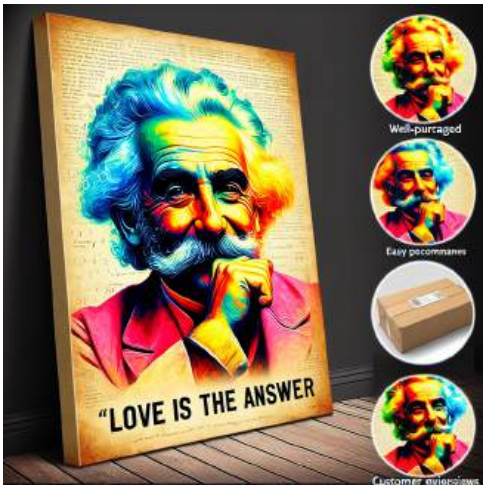


Figure 18. Dall-E Generated Image

Finally, this prompt explicitly asked for an image including Einstein and the sign with a message. While the model was able to generate an image with these elements, it did not accurately produce the sign in Einstein's hands. Additionally, the image included other physical characteristics retrieved from reviews, such as the product being well-packaged and easy to hang, which were irrelevant to this particular image.

## Discussion

In this RAG-Text-to-Image pipeline, the ability to generate images is heavily reliant on the availability and specificity of information in product descriptions and reviews. If the necessary details are missing or ambiguous, the model struggles to create accurate representations. It would be crucial to prioritize data preprocessing to ensure that textual inputs are detailed and unambiguous, providing sufficient information for the model to generate accurate images.

Despite advancements in AI, the generated images often exhibit unrealistic or "AI-like" characteristics. This includes incorrect scales (e.g., oversized luggage) and improbable scenarios (e.g., a coffee glass suspended in the air), leading to a lack of realism. Fine-tuning the model or using post-processing techniques can be considered to enhance the realism of generated images, focusing on aspects such as scale, perspective, and context.

The model tends to prioritize specific details from the prompts, potentially overlooking nuances and contextual information provided by broader language model inputs. For instance, while attempting to visualize a room, it might focus on one detail (e.g., making everything steel) while neglecting the overall context. One may consider implementing mechanisms to balance the model's attention between specific details and broader context, ensuring that it captures nuanced information effectively.

While AI text-to-image generation has made significant strides, our project highlights ongoing challenges related to realism, contextual understanding, and dependency on input data quality. Addressing these challenges will be crucial for further improving the practical applications of AI in visual content generation.