94-844: Generative AI Lab

Project Report

Ashwin Kandath Hoklam Cheung Praneet Yavagal Rahul Pujar Yukti Shah

Part 1: Product Selection

For this project, we selected three distinct products from different categories on Amazon, guided by revenue performance metrics across categories to ensure a representative sample of market dynamics. The categories include high-performing (Home and Kitchen), mid-range (Electronics), and low-performing (Musical Instruments) segments (Connolly, 2023). Specifically, we chose a coffee machine from Home and Kitchen, a category that consistently ranks as one of Amazon's top revenue generators due to its broad appeal, frequent purchases and also because it's a popular product with many similar variants. It is interesting to see how reviews are interpreted and reflected in image generation. For Electronics, we went with Meta Quest, a well-known and widely reviewed product, to explore how the system handles something familiar and mainstream, which might make it an easier task. Lastly, from the Musical Instruments category, we selected a handpan drum, a niche and relatively new product. We were curious to see how the system performs with something less common, where reviews might be more limited or specialized. These choices give us a mix of familiarity and challenge, helping us analyze how well the process works across different product types.

(L-R) Coffee machine, Meta Quest 3, Handpan Drum

Part 2: Analysis of Customer Reviews with LLM

In Part 2, we applied a paragraph chunking method, dividing the text into 37 chunks. Initially, we used the "text-embedding-3-small" model for retrieval, but the results were not entirely relevant to our queries. After switching to "text-embedding-ada-002," the retrieval quality improved significantly.

To enhance the system further, we implemented a keyword extractor to assist ChatGPT 4.0 in generating product descriptions. Additionally, we added a summary extractor to provide a more holistic understanding of the product context. We set K=5 to ensure the results contained minimal irrelevant information.

For the final output, we configured the ChatGPT 4.0 API to deliver a detailed description of the product's appearance, key features, and any potential failure points.

Prompt Generation Objective

Our goal in crafting these image-generation prompts was not to create a depiction identical to the actual product, but rather to leverage both positive and negative user reviews to craft a realistic and nuanced representation. By balancing praised features with constructive criticism, we aimed to ensure the visualizations conveyed an authentic and relatable product experience.

This approach serves several purposes:

1. Authenticity: It provides a more honest representation of the product, acknowledging both strengths and potential weaknesses.
2. User-Centric Perspective: By incorporating elements from user reviews, the image reflects real-world experiences rather than idealized marketing depictions.
3. Informed Decision-Making: Potential buyers can gain a more comprehensive understanding of the product, including possible limitations or areas of concern.
4. Balanced Representation: It avoids overly positive or negative bias, striving for a middle ground that respects diverse user experiences.
5. Visual Storytelling: By translating text reviews into visual elements, we create a more engaging and immediately understandable representation of user feedback.
6. Brand Trust: Demonstrating transparency about product strengths and weaknesses can potentially increase consumer trust in the brand.

The ultimate aim was to create prompts that would generate images offering valuable insights at a glance, helping consumers make more informed decisions while setting realistic expectations about the product.

Methodology

We implemented a Retrieval-Augmented Generation (RAG) system using the following components:

- Document Processing: PyMuPDF for PDF parsing
- Text Chunking: SentenceSplitter from llama_index. We applied a paragraph chunking method, dividing the text into 37 chunks.
- Embedding Model: OpenAI's "text-embedding-ada-002". We used the "text-embedding-3-small" model for retrieval, but the results were not entirely relevant to our queries. After switching to "text-embedding-ada-002," the retrieval quality improved significantly.
- Vector Store: Pinecone with cosine similarity metric
- Language Model: OpenAI's ChatGPT 4.0 API to deliver a detailed description of the product's appearance, key features, and any potential failure points.
- Keyword extractor to assist ChatGPT 4.0 in generating product descriptions.
- Summary extractor to provide a more holistic understanding of the product context. (We set K=5 to ensure the results contained minimal irrelevant information)

Experiments and Results

Breville Espresso Machines

1) Basic Summary

- The process began with a fundamental summary of Breville espresso machines, focusing on core features like design, functionality, and user experience.
- This summary aimed to provide a basic understanding of the product, which can serve as a foundation for creating a prompt for image generation.
- The initial prompt did not specify the need for visual details, and was more informational in nature, leaving out the elements necessary for image generation.

```
to_answer = "Give me a summary of Breville espresso machines"
to_use = "gpt-4"
answer = rag_openai_gpt(model_to_use, query_to_answer, combined_context)
```

2) Added a "Make it Visual" in the Prompt

- To enhance the image generation, the next step was adding the phrase "make it visual" to the prompt.
- This instruction directly communicates the need for DALL·E to focus on visual details when generating images.
- The idea was to ensure that the response would include specific visual elements like the appearance of the espresso machine, the design of its components, and the overall layout, helping to guide the AI in creating a more accurate and visually appealing image.
- The updated prompt provided DALL·E with a clearer focus on the visual aspects rather than just technical or functional details.

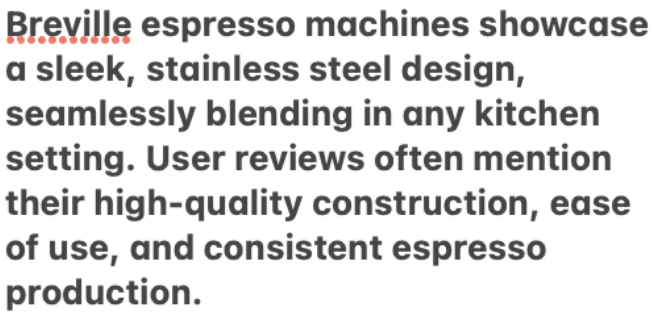
```
query_to_answer = "Provide a summary of Breville espresso machines, highlighting key features, design, \
                  and user benefits which are only useful for image generation"
model_to_use = "gpt-4"
final_answer = rag_openai_gpt(model_to_use, query_to_answer, combined_context)
```

3) Addition of User Benefits

- The next step involved incorporating user reviews and benefits to create a more user-centric perspective on the product. This inclusion allowed the prompt to consider how users perceive the espresso machine in real-life use.
- We expected that adding this context gave more richness to the prompt, especially when it came to representing how users experience the product's design and functionality.
- For instance, including benefits like "user-friendly controls" or "compact design" could guide DALL·E to consider how these aspects should be represented visually.
- User reviews also helped in focusing on practical, day-to-day interactions with the product, enhancing its realism and making the image generation more contextually accurate.

```
query_to_answer = "Give description of Breville espresso machines highlighting physical appearance \
                   and only visual aspects of user reviews in less than 20 words "
model_to_use = "gpt-4"
final_answer = rag_openai_gpt(model_to_use, query_to_answer, combined_context)
```

Output:



While adding user benefits was intended to provide richer context for the image generation, it did not fully deliver the expected visual depth in the prompt. Despite highlighting features like "user-friendly controls," "compact design," and "ease of use," the resulting image generation lacked the necessary visual richness to clearly depict these aspects. The benefits were primarily focused on functional and experiential aspects of the product, which are more abstract and harder for an AI model like DALL·E to translate into specific visual cues. For instance, terms like "user-friendly" or "ease of use" are subjective and don't offer the kind of concrete, visual details that a 3D artist would need, such as button layout, tactile feedback, or ergonomic design. As a result, the image generation did not fully capture the nuanced design features that directly relate to the user experience, missing the opportunity to create a more dynamic and visually representative image.

4) Addition of Context Saying "You Are Assisting a 3D Artist to Generate an Image"

- To provide even further clarity to the prompt, the next step was to explicitly mention that the task was to assist a 3D artist in generating an image.
- This instruction set the context that the generated images would be used for 3D modeling, where detail, perspective, and accuracy of the product's form and features are essential.
- By providing this direction, the prompt was aimed at ensuring that DALL·E would include technical details relevant to a 3D artist, such as visual depth, structure, textures, lighting, and the arrangement of objects.
- It also helped DALL·E understand that the output should be a high-quality representation with appropriate proportions and visual cues to suit a professional design process.

Output:

Adding context with the statement "You are assisting a 3D artist to generate an image. Give more details on how to get images exactly" significantly improved the outcome. The output became

much more focused on the essential visual details of the product, such as its "high-quality stainless steel build" and "user-friendly controls," which directly relate to the design and appearance of the espresso machine. These are tangible elements that a 3D artist could use to craft a visually accurate representation. Additionally, incorporating user feedback on features like "manual micro-foam milk texturing" and "customized coarseness of grinds" added more concrete, functional details that could be visually represented in the image, such as the grinder mechanism and the steam wand. This allowed for a clearer visual context, guiding the 3D artist toward specific areas of focus, like texture and functionality, making the prompt more actionable and relevant for image generation. Thus, adding the context did help by providing clearer directions to generate more visually rich and accurate images.

Breville espresso machines are robust and elegantly designed products known for their high-quality stainless steel built. Users praise them for consistent espresso production, user-friendly controls, and thoughtful design elements. Some users note high durability, easy usage, and impressive performance. A customized coarseness of grinds, manual micro-foam milk texturing, and easy dosage control are highly acclaimed features. However, few users indicate problems with grinder maintenance and concerns about premature failure. In summary, users find Breville espresso machines reliable and worthwhile for the exquisite coffee-making experience they provide.

Addressing Stable Diffusion Token Limit Challenge

Prompt creation for image generation in Stable Diffusion presented unique challenges due to the 77-token limit. This constraint required us to carefully distill extensive product descriptions, user reviews, and contextual details into concise yet meaningful prompts that could still convey visually rich and accurate outputs.

The Issue: Token Limit

Stable Diffusion processes text prompts with a maximum input size of 77 tokens. Anything exceeding this limit gets truncated, leading to incomplete or inaccurate image generation. This posed a problem when trying to encapsulate comprehensive product details, such as:

- Design attributes (materials, shapes, colors).
- Functionality highlights (key features, use cases).
- Contextual elements (user reviews, benefits, and critiques).

Since user reviews often include nuanced feedback about products, compressing this information without losing its essence required strategic rewriting and prioritization.

Our Approach

1. **Iterative Summarization**
    - We broke down long product descriptions and user reviews into their core components: essential features, pros, and cons.
    - Each component was then rewritten in a concise, descriptive format, ensuring it could fit within the token limit.
2. **Focus on Visual Specificity**
    - To align with Stable Diffusion's strengths, the prompts emphasized visual details. For example, instead of saying *"the machine is compact and user-friendly,"* we used *"a sleek stainless-steel espresso machine with a compact design, user-friendly controls, and a built-in grinder."*
3. **Iterative Testing and Refinement**
    - Each version of the prompt was tested in Stable Diffusion to evaluate the quality of generated images. Adjustments were made to prioritize the most visually impactful details, ensuring each prompt delivered a clear and realistic image.
4. **Balancing Positives and Negatives**
    - User reviews (both positive and negative) were incorporated subtly to create prompts that highlighted realism. For instance, references to *"robust design with occasional grinder maintenance issues"* were included to add authenticity to the generated images.

Final Prompts

The final prompts for the three products were carefully crafted to:

- Stay under 300 characters to avoid truncation.
- Include detailed, visual descriptions of product features.
- Reflect user feedback to make the images more realistic and relatable.

These finalized prompts have been stored in the **deliverables folder** for future use for each of the 3 products:

<u>Breville Espresso Machines</u>

Sleek Breville espresso machine with stainless steel body, built-in grinder, and LCD display. Steam wand frothing milk into intricate latte art. Pressure gauge glowing subtly. Nearby, a perfect espresso shot emits rich crema. Icons hover: a glowing thumbs-up with vibrant coffee beans and a wrench with faint cracks. Water tank half-full, reflecting soft light.

Meta Quest 3

Meta Quest 3 512GB VR headset with controllers. 4K display emits vibrant holograms. Floating icons: 5-star badge with thumbs-up, smiling avatar in mixed reality; 3-star badge with depleting battery, frowning avatar. Batman logo and Quest+ trial emblem hover nearby. Background shows silhouette seamlessly interacting with virtual objects. High-quality carrying case beside headset. Faint '2064x2208' text per lens.

Handpan Drum

Shimmering handpan drum with 9 circular indentations, mallets resting on its surface. Floating icons: golden musical note radiating concentric waves, silver tuning fork with jagged lines. Transparent hands playing drum: one emits colorful notes, other produces dull, faded tones. Background: intertwined musical staff with crisp and blurry notes. Sturdy travel case contrasts with torn fabric bag. Lotus flower and om symbol hover above.

Key observations for prompt design:

1. Specificity: Each prompt explicitly requested visual details, key features, and user experiences. This specificity guided the RAG system to retrieve and synthesize relevant information.
2. Focus on image generation: By including "for realistic image generation" in each prompt, the system was directed to prioritize visually descriptive information.
3. Multi-faceted approach: The prompts covered multiple aspects (appearance, features, user interaction) to ensure a comprehensive description.
4. Product-specific elements: Each prompt was customized to the product type, e.g., "playing technique" for the handpan drum, "user interaction" for Meta Quest 3.
5. Conciseness: The prompts were designed to be clear and concise, typically consisting of a single sentence.

Parameter Experimentation

1. Vector Store Query (top_k):
    - Tested values: 5-15
    - Observation: Higher values provided more context but increased noise.
    - Optimal value: 10 for a balance of information breadth and relevance.
2. Embedding Model:
    - Used: "text-embedding-ada-002"
    - Finding: Consistent performance across diverse product categories.
3. Language Model Configuration:
    - Used: GPT-4o

- Insight: Allowed for more nuanced prompts and better handling of complex tasks compared to GPT-3.5-turbo.

4. Prompt Structure:
    - Specific instructions tailored to each product category resulted in targeted responses.
5. Token Management:
    - Implemented a 2000 token limit for context retrieval.
    - Benefit: Balanced context retrieval without overwhelming the model

Key Observations for Parameters

1. Retrieval Effectiveness: The system consistently retrieved relevant information across diverse product categories, demonstrating robustness in the embedding and vector search approach.
2. Context Integration: Successfully combined multiple text chunks to provide comprehensive answers within a 2000-token limit.
3. Response Quality: Generated responses were coherent and directly addressed queries, indicating effective integration with the language model.
4. Versatility: Handled queries about different product types equally well, showcasing adaptability.

Part 3: Image Generation with Diffusion Model

In this part, we evaluated the ability of diffusion-based AI models to generate accurate and visually appealing product images from textual descriptions and customer reviews. Using insights extracted in Part 2, we crafted prompts and iteratively refined them to guide image generation. Our aim was to assess how closely AI-generated images matched real-world product images and to compare the outputs of two leading models: DALL-E 3 and Stable Diffusion 3.5 Large.

Image Generation Models Used:

- **DALL-E 3:**
    - Known for artistic flair and flexibility.
    - We used it to explore creative interpretations of product visuals.

- **Stable Diffusion 3.5 Large:**
    - Specialized in realistic and precise visual depictions.
    - Prioritized for its ability to adhere closely to prompts.

Prompt Design:

Prompts were derived from the outputs of Part 2. Each JSON file contained key visual details extracted from product descriptions and customer reviews, which were formatted into a standardized prompt template.

- Product Descriptions: Key specifications and characteristics were identified, including size, shape, color, and material.
- Customer Reviews: Feedback on design, performance, and usability helped enrich the visual details of the prompts.

Example JSON Output from part 2 (Breville Coffee Machine):

{
  "prompt": "Sleek Breville espresso machine with stainless steel body, built-in grinder, and LCD display. Steam wand frothing milk into intricate latte art. Pressure gauge glowing subtly. Nearby, a perfect espresso shot emits rich crema. Icons hover: a glowing thumbs-up with vibrant coffee beans and a wrench with faint cracks. Water tank half-full, reflecting soft light."
}

Example General Prompt:

```
# Extract the image prompt
image_prompt = prompt_data.get('prompt')
final_prompt =(
    f"A high-quality, professional product photo of {image_prompt}.\n"
    f"The image should accurately represent the product's visual details.\n"
    f"Style: studio lighting, plain background, **no text**, **no logos**, **no text labels**, **no watermarks**.\n"
    f"Focus solely on the product."
)
```

Note: The image prompt here comes from the json output of part 2 as shown above.

Results:

1. Breville Coffee Machine

Description from Part 2:

- Sleek Breville espresso machine with stainless steel body, built-in grinder, and LCD display. Steam wand frothing milk into intricate latte art. Pressure gauge glowing subtly. Nearby, a perfect espresso shot emits rich crema. Icons hover: a glowing thumbs-up with vibrant coffee beans and a wrench with faint cracks. Water tank half-full, reflecting soft light.

Original Image:

Generated Images:

- DALL-E 3:
    - Successfully captured the stainless steel finish, creating an appealing, polished look.
    - Added creative lighting effects that emphasized the product's modern and premium feel.
    - Missed some finer design details, like the exact layout of the buttons and the proportions of the grinder.

- Stable Diffusion:
    - Delivered realistic visuals closely aligned with the actual product’s dimensions and features.
    - The grinder, LCD display, and overall shape were accurately rendered. However, the lighting and shadows were more subdued, which slightly diminished the premium feel compared to DALL-E.

Comparison with Real Product:

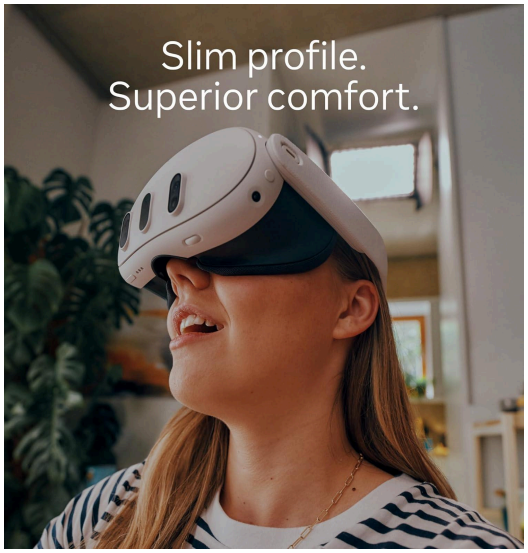
- While DALL-E improves visual appeal but deviates significantly from realism, Stable Diffusion's outputs accurately match the durability and accuracy of the real product. The fine elements of the design, such as the steam wand's subtle curve and the grinder's exact texture, were difficult for both models to accurately capture.

2. Meta Quest 3

Description from Part 2:

- Meta Quest 3 512GB VR headset with controllers. 4K display emits vibrant holograms. Floating icons: 5-star badge with thumbs-up, smiling avatar in mixed reality; 3-star badge with depleting battery, frowning avatar. Batman logo and Quest+ trial emblem hover nearby. Background shows silhouette seamlessly interacting with virtual objects. High-quality carrying case beside headset. Faint '2064x2208' text per lens.

Original Images:



Generated Images:

- DALL-E 3:
    - Created futuristic and aesthetically pleasing depictions, emphasizing the headset’s technological nature with dramatic lighting and smooth curves.
    - Sometimes exaggerated features, such as over-rounded edges or overly bright highlights, which made the images slightly less accurate.

- Stable Diffusion:
    - Produced images that were nearly identical to the real product in terms of shape and color.
    - The headset's ergonomic design and sharp resolution cameras were rendered with a high degree of accuracy.
    - The lighting was natural but less dynamic compared to DALL-E's creative interpretations.

Comparison with Real Product:

- The real product's minimalistic elegance is more closely captured by Stable Diffusion, which adheres to precise proportions and textures. DALL-E introduces an artistic flair that enhances appeal for marketing purposes but may mislead consumers expecting an exact match.

3. Handpan Drum

Description from Part 2:

- Shimmering handpan drum with 9 circular indentations, mallets resting on its surface. Floating icons: golden musical note radiating concentric waves, silver tuning fork with jagged lines. Transparent hands playing drum: one emits colorful notes, other produces dull, faded tones. Background: intertwined musical staff with crisp and blurry notes. Sturdy travel case contrasts with torn fabric bag. Lotus flower and om symbol hover above.

Original Images:

Generated Images:

- DALL-E 3:
    - Added intricate patterns and textures not present in the actual product, such as stylized etchings or exaggerated indentations.
    - Captured the drum's golden tone but sometimes introduced color variations, making it appear more bronze or metallic gray.

- Stable Diffusion:
    - Rendered realistic representations of the drum's size, shape, and material.
    - Maintained the simple yet elegant design without introducing unnecessary artistic elements.
    - Occasionally underplayed the reflective quality of the golden surface, making it look slightly duller than the real product.

Comparison with Real Product:

- The real product's elegant simplicity is better mirrored by Stable Diffusion, while DALL-E's artistic additions provide a creative reinterpretation. For niche items like the handpan drum, Stable Diffusion's focus on accuracy made its outputs more relatable to the actual product.

Strengths and Challenges of Each Model:

1. **DALL-E 3**:
    - **Strengths**:
        - Adds a dynamic, artistic touch, making outputs visually engaging.
        - Especially effective for products like the Meta Quest 3, where a futuristic feel enhances the appeal.
    - **Challenges**:
        - Often strays from strict realism by introducing artistic embellishments.
        - Struggles to capture intricate design details accurately, as seen with the Breville Coffee Machine.

2. **Stable Diffusion 3.5**:
    - **Strengths**:
        - Delivers realistic outputs with a strong focus on adhering to product descriptions.
        - Especially reliable for simpler designs like the handpan drum, where its straightforward approach excels.
    - **Challenges**:
        - Lacks the dramatic lighting and artistic enhancements that make products visually striking.
        - Occasionally misses nuanced textures and reflective details, as observed in the Breville Coffee Machine.

Comparison of Models based on specific criteria:

Criteria	DALL-E	Stable Diffusion
Aesthetics	Superior control with professional-grade lighting effects	Natural shadow placements with some inconsistencies
Accuracy	Vivid and precise color reproduction enhancing product appeal	Generally accurate with minor color deviations in certain variants
Consistency	Rich, detailed backgrounds that may sometimes distract from the product	Clean, minimalistic backgrounds that emphasize the product
Detailing	Addition of unwanted details, slight deviations from expected outcomes	Excellent at rendering fine details and sticking to the prompt

Conclusion:

In conclusion, our analysis revealed that DALL-E and Stable Diffusion have distinct strengths that cater to different use cases in text-to-image generation. DALL-E excelled in creating visually creative and artistically appealing outputs, making it ideal for marketing or promotional materials where aesthetics are a priority. Conversely, Stable Diffusion demonstrated superior accuracy and realism, aligning closely with product specifications, making it better suited for e-commerce applications or scenarios where precise depictions are essential. Both models showcased significant potential but required iterative refinement of prompts to achieve optimal results.

We learned that effective prompt design, informed by detailed product descriptions and customer reviews, is critical to achieving accurate and meaningful visualizations. Iterative testing played a pivotal role in improving output quality, emphasizing the importance of a dynamic approach to leveraging these AI tools.

Thus, hybrid workflows that combine the creativity of AI models with human oversight could further enhance the effectiveness and reliability of text-to-image generation. Additionally, fine-tuning models for specific product categories or domains could address challenges in handling less familiar items, opening avenues for more tailored and impactful applications of generative AI.

References

December 6, Brian Connolly, and 2023. 2023. “Top Amazon Product Categories in 2024.” Jungle Scout. December 6, 2023. https://www.junglescout.com/resources/articles/amazon-product-categories/.

“Stabilityai/Stable-Diffusion-3.5-Large · Hugging Face.” 2024. Huggingface.co. October 29, 2024. https://huggingface.co/stabilityai/stable-diffusion-3.5-large.

Appendix

Products

Coffee Machine: Amazon.com: Breville BES870XL Espresso Machine, One Size, Brushed Stainless Steel

Meta Quest 3: Meta Quest 3 512GB — The Most Powerful Quest — Ultimate Mixed Reality Experiences — Get Batman: Arkham Shadow and a 3-Month Trial of Meta Quest+ Included

Handpan Drum: Amazon.com: Handpan Drum 22 Inches D Minor Kurd, 432Hz 10 Notes Hand Drum Instrument, Premium Steel Hand Drum, Handpan Instrument Handpan Drum for Adults (Gold)