SephoraShopper: Personalized Product Review Generation

Hilari Fan
Univ of Southern California
hilarifa@usc.edu

Wonjun Lee
Univ of Southern California
wonjunle@usc.edu

Seena Pourzand
Univ of Southern California
spourzan@usc.edu

Abstract

With an extensive catalog of beauty products, Sephora can be an intimidating experience for online shoppers. We propose SephoraShopper, a natural language processing tool that simplifies the decision-making process. By allowing shoppers to input their personal characteristics and a product of interest, SephoraShopper generates a predictive review that anticipates what the user would write after using the product. Our solution aims to transform the online cosmetic shopping experience, presenting users with a personalized preview to help them make informed purchasing decisions.

1 Introduction

The beauty industry has become an increasingly saturated market. Projected to reach around $580 billion by 2028, this industry is experiencing unprecedented growth. While impressive, this growth presents a significant challenge for consumers navigating this saturated space. Consumers are increasingly overwhelmed by the sheer volume of cosmetic products on the market and experience a phenomenon known as "choice overload." This problem is especially evident in Sephora, the leading omni-retailer with over 2,700 stores worldwide and a growing online presence.

Sephora's product range includes thousands of items, each with up to 17k reviews. With the e-commerce growth in the industry, more consumers are purchasing cosmetics online but the volume of products and reviews makes it difficult for consumers to choose the best product for their unique needs.

SephoraShopper leverages LLM to generate personalized product reviews. Unlike traditional review summaries, SephoraShopper predicts the content of a review that a user is likely to write based on their specific characteristics, the product details, and existing reviews. This tool simplifies the product selection process by providing insights into how well a product aligns with a user's preferences and characteristics. By simulating personalized reviews, SephoraShopper enables consumers to make informed decisions among the countless beauty products available.

2 Problem within Related Work

Various previous works explore the use of natural language generation (NLG) specifically for tasks involving product reviews using transformer models.

Dong et. al., 2017 explored an attention-enhanced approach to generate product reviews when conditioned on user, product, and rating attributes. They used multi-layer perceptrons to encode user and item IDs into context vectors. Then, they decoded a word sequence, which is the generated reviews, using stacked recurrent neural networks and introduced an attention mechanism to learn the association between the context vector and the predicted output words. They utilized an Amazon book product dataset and demonstrated that their model outperforms baseline methods, such as nearest neighbor search, by leveraging the attention mechanism. This research has laid a foundation for generating personalized review content based on user and product attributes.

Li et al., 2020 examine the possibilities of personalized NLG with their model, PErsonalized Transformer for Explainable Recommendation (PETER). Similar to Dong et al.'s work, they used user and item-specific information for content generation. However, Li et al. pioneered the use of transformer models in this domain, demonstrating that PETER can both generate explanations and make recommendations based on these attributes. They introduced an innovative approach that integrated user and item IDs with textual data and revised the attention-masking matrix to accommodate their

three tasks: explanation generation (creating text rationale behind a product recommendation to a user), context prediction (mapping user and item IDs to words used in the explanation), and rating prediction (estimating a user's potential rating based on historical data and user/product attributes). PETER outperformed fine-tuned BERT models in several metrics, underscoring the potential of transformer models for personalized NLG tasks that require an understanding of user and item-specific information.

Like Dong et al., we attempt to tackle the task of generating product reviews. Unlike their research which focuses on attribute-to-sequence generation with RNNs, our SephoraShopper draws inspiration from Li et al., 2020 by leveraging the transformer model for personalized NLG. SephoraShopper attempts to generate personalized product reviews that reflect the unique characteristics of individual users and product details by utilizing GPT-2, T5, Llama, and Mistral.

3 Hypothesis

We hypothesize that a transformer model, fine-tuned on Sephora product reviews and product descriptions, can accurately predict the content that a user with a specific set of characteristics (e.g. skin type, skin tone, product detail desires) is likely to write.

Transformer models are highly effective for tasks involving text generation, making them an excellent choice for our Sephora personalized product review generation use-case. Transformer models excel in text generation task because of the attention mechanism which allows the model to capture the contexts of the input sequence and their contribution to the predicted output. In addition, we leverage pre-trained transformer models like GPT-2, T5, Llama, and Mistral fine-tuned to our dataset of Sephora product reviews and details to create personalized product reviews that can guide the user's purchasing decisions and enhance their online shopping experience.

4 Dataset

We collected a comprehensive dataset of over 40,000 product reviews, scraped from the Sephora website. We began our data collection process by strategically selecting the products to include in our dataset. We identified fifteen foundation products for analysis, prioritizing diversity in product characteristics, such as suitability for different skin types (oily, dry, or mature), finish (natural, satin, matte), and coverage (light, medium, full). We chose foundation products because individual user characteristics and preferences significantly influence a consumer's choice in complexion products as compared to other cosmetic categories.

Initially, we attempted to gather the data by using traditional web-scraping techniques using Beautiful Soup. However, we transitioned to using API calls to the Bazaarvoice network, which hosted all of the review and product info data. This approach allowed us to retrieve product information, which includes product ingredients, finish, coverage, and description, and the review information, which includes review text, rating, recommendation status, positive and negative feedback counts, skin tone, skin type, presence of incentivized review, and verified purchaser status.

Once we collected the data, we preprocessed the data. To address missing values, we marked them as "empty" and filled in missing data points for incentivized review and verified purchaser fields to ensure the completeness of our datasest. We also excluded the reviews that were incentivized to minimize the potential bias that these reviews could introduce. Furthermore, we cleaned the dataset by trimming excess whitespace to standardize the data format.

From data scraping, we obtained a CSV of data, with each datapoint or user/product detail in a separate column. We then prepared a CSV that contained a singular input and output column, a format that makes it easy to extract input-output pairs for the models. Each entry of our original CSV was taken row by row and concatenated into an input string for the product and user details while the review and rating were concatenated to create the output. A sample input and output entry is provided in Table 2.

5 Approach

5.1 Generative Pre-trained Transformer 2 (GPT-2)

GPT-2, developed by OpenAI, has shown a significantly advanced ability to generate text. The model has a decoder-only architecture which has an advantage of simplicity and faster training speed over models with an encoder-decoder archicture. GPT-2 is pre-trained on a corpus called WebText with approximately 8 million web pages, making

the model well-suited for understanding our review and product text input and generating personalized predictive reviews.

We further preprocessed the data by combining the input and the output into a singular string, which we wrote to a text file. This approach exposes the model to both the input and output text during training to learn to generate appropriate product reviews given an input. We used the GPT-2 tokenizer to encode input-output pairs, the input as all of the product and user information and the output as the review text and rating. We loaded the tokenized data into a TextDataset class and trained our model using GPT2LMHeadModel within Hugging Face's Transformers library.

We fine-tuned the model by adopting a learning rate of 2e-5, a batch size of 32 for training and 64 for evaluation, and eval steps of 500 over 3 epochs.

5.2 Text-to-Text Transfer Transformer (T5)

The T5 model, developed by Google, is pre-trained on a large corpus of unsupervised and supervised tasks, including text-to-text tasks such as sentiment analysis, question answering, and natural language inference. T5 has an encoder-decoder architecture, which offers an advantage over decoder-only models like GPT-2 because this structure allows a deeper, more complex contextual understanding of the input. Given the lengthy, information-dense input sequences containing user information and long product descriptions, this is crucial for our review generation task.

We prepared our dataset into Prompt-Response pairs by converting our preprocessed CSV data into input and output pairs. To guide the model on our task, we prepended the prefix "generate review: " to each input text to clarify the task context to the model.

We tokenized the input and output data using the T5 tokenizer and used the T5ForConditionalGeneration class for our model, which is tailored for conditional generation tasks like ours. Furthermore, the T5Tokenizer utilizes SentencePiece which combines subword units from Byte Pair Encoding and treats text as a raw input stream to be agnostic to spaces for other languages. Although that it isnt particularly relevant since our project deals with English, SentencePiece is also used in conjunction with Unigram which begins at a base vocabulary that is trimmed when computing how the loss changes after trimming. These two approaches that work opposite to each other, merge subwords and trimming base words help give a very diverse tokenization approach. Given the computational constraints, we selected the t5-small, the most lightweight version of the T5 models. We trained the T5 model and fine-tuned the following hyperparameters: 3 epochs, 8 train and test batch size, 500 eval steps, 2e-5 learning rate, and 0.01 weight decay.

5.3 Llama 2

Llama 2, developed by Meta, is an auto-regressive language model with a decoder-only architecture that adopts grouped-query attention to optimize the generation of human-like text (Ainslie et al., 2023). Meta has previously fine-tuned this model into Llama 2 Chat for dialogue use cases, incorporating supervised fine-tuning and reinforcement learning with human feedback (RLHF) (Touvron et al., 2023). Llama 2 is a substantially larger language model than both GPT-2 and T5, with 7 billion parameters compared to GPT-2's 1.5 billion and T5-small's 60 million. It was trained on 2 trillion tokens, significantly more than GPT-2's 50,257 and T5's 1 trillion tokens. We selected this model to explore whether a larger number of parameters could outperform earlier models with our Sephora dataset. In addition, Llama 2's experimental results may clarify the benefits of T5's encoder-decoder architecture versus the decoder-only architectures of GPT-2 and Llama 2 when handling our dataset that has clear input and output distinction.

We followed a specific prompt, predefined by the model, in order to fine-tune more efficiently. The training input was prepended with start tokens <s> and enclosed with instruction tokens [INST] and [/INST]. </s> was appended to both the input and output to signal the end of data. An example looks as follows: "<s>[INST] input [/INST] output </s>". Without these tokens, the model performed poorly, generating irrelevant text to the product or reviews.

The tokenizer was LlamaTokenizer, which uses byte-pair encoding based on a SentencePiece. The tokenizer does not automatically append a prefix to the first token, thus necessitating the <s> token.

Our fine-tuning also involved a QLoRa configuration, which combines quantization and Low-Rank Adapter (LoRA) to enable Parameter-Efficient Fine-Tuning (PEFT) (Dettmers et al., 2023). Given Llama 2's significantly large size of 7 billion parameters, the GPU was unable to manage such demands. Therefore, PEFT was essential to

reduce the number of parameters. The 4-bit quantization subdivides the pre-trained Llama 2 model into 4 bits and keeps their parameter frozen. A smaller-sized LoRA layer, whose parameters are not frozen and are updated during fine-tuning, is added. During training, the model only tracks gradients for the backpropagation of the LoRA layers, not the frozen 4-bit models, therefore, significantly saving space and computational time by updating a smaller subset of 7 billion parameters and circumventing the problem of limited GPU resources.

The hyperparameter used for Llama 2 was 1 epoch, training batch size of 1, learning rate of 0.0002, and adam optimizer. Despite PEFT with QLoRA configuration, the GPU resource could only handle batch size of 1 (anything greater would result in out of memory error). Llama 2 used higher learning rate and optimizer of Adam, which is different from the previous two models. This is because the model needed to converge faster in shorter amount of time (i.e. learn more in less epoch). High learning rate updates the parameters faster to meet its optimal weights while the optimizer helps converge faster and more accurately.

5.4 Mistral

We opted to utilize Mistral 7B Instruct v0.2, the instruction tuned version of Mistral 7B with no sliding window attention. The main motivation to use Mistral is to use another decoder-only model to better differentiate between the performance of T5's encoder-decoder model. Mistral also utilizes group query attention to help reduce cache size and increase inference speed along with its 32k context window. Mistral uses Byte Pair Encoding similar to GPT2 and SentencePiece(similar to T5) as well, both methods building subwords albeit using different approaches. According to Mistral AI's blogpost, Mistral 7B model outperforms Llama 13B at text generation tasks, making it a cost efficient yet effective option for our research problem, only looking at the attention of a group of hidden states instead of all. For our hyperparameters, given our computational resources, we had a batch size of 4, weight decay 0.001, a learning rate of 23-4, 32bit adam for our optimizer, and half an epoch for training with floating point 16 bit representation enabled to help reduce size.

6 Evaluation

To evaluate the performance of our models, we used three common performance metrics: Bidirectional Encoder Representations from Transformers Score (BERTScore), BiLingual Evaluation Understudy (BLEU), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE). The BERTScore evaluates the semantic similarity between the generated review text and the reference user reviews. It leverages the contextual embeddings from pre-trained models like BERT, comparing the similarity of words in the generated text to those in the reference text. This metric goes beyond word matches or n-gram overlap, making it particularly relevant for our task. The BLEU score measures the precision of how well the generated reviews match the reference reviews in terms of word choice overlap and sentence structure. Although commonly used in translation tasks, it could help us assess the accuracy of the model's generated review text in replicating the style and content of the Sephora reference reviews. We use the ROUGE score, specifically ROUGE-1 and ROUGE-2, to measures the recall, which is the proportion of reference text instances captured by our generated model text. This metric is typically used in machine summarization tasks and helps evaluate the extent that key information from the reference reviews is retained in the generated review.

For each generated product review, we calculate the BERTScore, BLEU, and ROUGE scores and compare these against a set of reference reviews for that product. Together, these metrics provide a comprehensive assessment of the performance of our model in generating contextually appropriate and accurate reviews.

7 Results

A sample generated output can be seen in Table 2. We evaluated the performance of four models: GPT-2, T5, Llama 2, and Mistral using the three metrics: BERTScore, BLEU, and ROUGE-1, as summarized in Table 1.

GPT-2 had the lowest BLEU score or 0.0022 indicating minimal word matches with the referenced reviews. It achieved moderate precision, recall, and F1 in BERTScore; however, the ROUGE-1 scores were also relatively low. A qualitative analysis on the generated output text from Table 2 suggests that GPT-2 generates a review that is somewhat repetitive and illogical. The text claims the foundation

Model	BLEU	BERTScore	ROUGE-1
GPT-2	0.0022	P: 0.788, R: 0.839, F1: 0.812	P: 0.090, R: 0.259, F1: 0.126
T5	**0.0524**	**P: 0.864, R: 0.889, F1: 0.876**	**P: 0.434**, R: 0.367, **F1: 0.373**
Llama 2	0.0077	P: 0.771, R: 0.869, F1: 0.817	P: 0.072, **R: 0.443**, F1: 0.121
Mistral	0.0003	P: 0.763, R: 0.843, F1: 0.801	P: 0.102, R: 0.395, F1: 0.154

Table 1: Comparison of NLP Models

has a "medium coverage" and contradicts this in the following line with a "light coverage." Similarly, the review contains some negative feedback, followed by high remarks and a 5-star rating.

In contrast, the T5 model achieved the best performance across all of the metrics with a BLEU score of 0.0524, BERTSCORE with the highest precision (0.864), recall (0.889), and F1 (0.876), and ROUGE-1 score with the highest precision (0.434) and F1 (0.373). This suggests that T5 generates product reviews with a higher semantic similarity to, precision, and recall from the reference reviews. Since the other three models have a decoder-only architecture, this suggests that the reason T5 performs better may be attributed to its encoder-decoder architecture which allows it to learn the complex relationships within the input. The generated sample T5 review is contextually accurate and aligns closely with the input details, such as the user's oily skin type.

Llama 2 has a moderately low BLEU, BERTScore, and ROUGE-1 scores, with a BERTScore F1 score (0.817) and ROUGE-1 F1 score (0.121) comparable to GPT-2. We trained Llama 2 for only 1 epoch due to GPU resource limitations, so the reduced training time could contribute to the observed lower results. The sample Llama 2 review is more logical and coherent than the GPT-2 output, but fails to capture some of the input details, writing that the user has a dry skin type.

Finally, Mistral obtained the lowest BLEU score (.0003) and BERTScore and moderately low ROUGE-1 scores. This suggests that Mistral is poor at semantic matching and capturing the context of reference reviews. Similar to Llama 2, Mistral is a significantly larger model than GPT-2 and T5, which despite their potential for higher performance, presented practical challenges during fine-tuning with our limited computational resources available. Like Llama we only trained the model for 1 epoch. The generated Mistral output is rich in detail and length, although some specificity seems randomly generated rather than inferred from the input context.

8 Conclusion

Among the various models we tested, we found that T5 performed the best. To better determine the reason for its performance, we attempted to look at more Decoder-only architectures to narrow down what is responsible for its superior performance. Mistral and Llama 2 are both SOTA of Decoder-only LLM's with billions of parameters. However, it is worth noting the sheer size of these SOTA Decoder Only models made it difficult to train meaningfully. Even with quantization efforts like PEFT with QLoRA, training was restricted to only one epoch with very small batch sizes without maxing out the available GPU and compute. T5 on the other hand comes with various sizes that make using it much easier given these constraints. With these smaller sized variants, we were still able to train without the need for quantization and for a decent amount of epochs as compared to LLMs. Furthermore, we suspect its Encoder-Decoder architecture that focuses on sequence to sequence transformations allows for it to develop and much more rich and indepth semantic understanding of the text, as not only is text generation a task but machine translation and summarization for example are also a part of the training for the model.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models.

Achim Berg, Sara Hudson, Kristi Klitsch Weaver, Megan Lesko Pacchia, and Imran Amed. 2023. The beauty market in 2023: A special state of fashion report.

Context (Input)	Skin Type: oily | Skin Tone: fair | Verified Purchaser: False | Positive Feedback: 0 | Negative Feedback: 0 | Recommended: True | Product: What it is: A hydrating foundation that delivers buildable coverage for the face and body, resists heat and humidity, and leaves a luminous makeup look. Coverage: Medium Finish: Natural Formulation: Liquid Highlighted Ingredients: - Squalane- Hyaluronic Acid What Else You Need to Know: This foundation is formulated with 94 percent natural-origin ingredients and hyaluronic acid. It delivers intense hydration and a second-skin sensation, while providing buildable coverage that evens the skin and blurs imperfections for a wide range of effects, from a flawless no-makeup look to a high-perfection complexion.
Ground Truth Output	Review: New favorite foundation. Easy to blend | Rating: 5
GPT-2 Output	Review: I love this for dry skin as this is so easy to use. I do usually wear a foundation brush to apply and get nice medium coverage. This makeup also does last a long time so I'm loving the light coverage. The only drawback is it blends out too long and is hard to blend out the foundation, but I think that's the foundation! Definitely a must-have! | Rating: 5
T5 output	Review: I love this foundation. It's buildable and looks great on my skin. It's a great foundation for oily skin. | Rating: 5
Llama 2 Output	Review: I love this product! It's my go to everyday foundation. It has great coverage and lasts all day long without any break out or caking up. The color match was perfect as well. I have dry skin so I need something moisturizing but not too heavy. This does both perfectly. | Rating: 5
Mistral Output	Review: I love this foundation! It's lightweight, buildable, and has a natural finish. I have combination skin and this foundation doesn't make me look oily or dry. It's also very long-lasting and doesn't break down in my t-zone. I have a lot of redness and this foundation covers it without looking cakey. I have a lot of texture on my cheeks and this foundation doesn't settle into it. I have a lot of pores on my nose and this foundation doesn't make them look bigger. I have a lot of fine lines around my eyes and this foundation doesn't make them look more pronounced. | Rating: 5

Table 2: Generated output based on the input to GPT-2, T5, Llama 2 after fine-tuning

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. pages 623–632, Valencia, Spain. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. pages 623–632. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

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.