ReviewRefine: Enhancing Transparency in E-commerce and evaluating AI Summarization Techniques for Amazon Reviews

Adeline Liou
adeline@usc.edu

Tais Mertz
tmertz@usc.edu

Abstract

The exponential growth of e-commerce, notably accelerated by the pandemic, has positioned Amazon as a dominant force in online retail. However, navigating the vast array of products in a straightforward way is difficult for consumers. Amazon has responded by leveraging AI and natural language processing (NLP) to streamline user experience, particularly in product discovery and review analysis. ReviewRefine aims to use AI summarization models like bart-large-cnn and flan-t5 to improve the user experience and improve transparency on the Amazon marketplace. By comparing these summaries with existing sources of information and evaluating the sentiment and similarity between these components, the study aims to assess how models like bart-large-cnn and flan-t5 have the potential to enhance transparency in e-commerce, ensuring consumers receive genuine representations of product sentiments in the complex online marketplace.

1 Introduction

E-commerce has experienced unprecedented growth in recent years, particularly accelerated by the pandemic, with Amazon emerging as a leader in the online shopping realm. This rise has been fueled by a cross-side positive network effect, attracting an ever-increasing number of both consumers and sellers. As the marketplace burgeons with a diverse range of products, consumers face the daunting task of navigating through an abundance of choices without the tactile advantages of traditional shopping—such as physically assessing product size, feel, and quality.

Amazon has made strides to enhance user experience and manage this complexity by deploying AI and NLP techniques. One notable feature involves aggregating similar products across various brands to aid consumers in making informed decisions based on comparative descriptions. However, the reliability of these seller-provided descriptions often remains questionable, reflecting an underlying issue of transparency in how products are presented online.

To further assist consumers, Amazon has also capitalized on the expansive pool of product reviews—an invaluable resource for prospective buyers. Notably, while 90% of consumers consult these reviews before making a purchase, the sheer volume and diversity of opinions can be overwhelming and time-consuming to sift through. Additionally, the authenticity of these reviews is frequently compromised by the proliferation of manipulated or fake reviews aimed at enhancing seller reputations artificially.

In response, Amazon has implemented large language models to summarize these reviews, aiming to distill the overall sentiment into a concise, digestible format. This summarization attempts to balance the positives and negatives mentioned in customer feedback, providing a structured snapshot of user experiences. However, the effectiveness and accuracy of these summaries in reflecting the true content of extensive user reviews remains an area ripe for exploration.

Our research aims to bridge these two facets—review transparency and summarization efficacy—by evaluating how well models like bart-large-cnn and FLAN-T5 perform in summarizing Amazon reviews accurately and reflectively. By looking at the performance of these models and comparing the results with the existing sources of information, we seek to understand their potential to enhance transparency in e-commerce settings, ensuring that consumers receive a true representation of product sentiments as they navigate the complex online marketplace.

1.1 Purpose

The purpose of this research is to critically evaluate the efficacy of advanced natural language processing (NLP) models, such as bart-large-cnn and FLAN-T5, in summarizing product reviews on Amazon. Our study aims to address several pervasive issues in e-commerce that compromise consumer decision-making: the overwhelming volume and diversity of product reviews, the prevalence of fake reviews, and inherent biases in consumer feedback. These factors contribute to a lack of transparency in online product presentation, with Amazon and sellers often having vested interests in promoting sales.

By assessing how well these NLP models summarize user reviews, we seek to determine their ability to provide a balanced and accurate reflection of the true content and sentiments expressed by consumers. This investigation is particularly crucial in understanding

whether these technological solutions can effectively cut through the noise of mixed reviews, counteract the effects of fake reviews, and mitigate biases that may skew consumer perception. Ultimately, our research intends to contribute to enhancing transparency in e-commerce platforms, ensuring that consumers have access to reliable and concise summaries of product reviews that aid them in making informed purchasing decisions.

1.2 Existing Work

Several products and tools have been developed to aid consumers in navigating the vast and often misleading world of online reviews. Our research draws upon the foundations set by these existing technologies, examining their methodologies and shortcomings to better understand how our work can contribute to this evolving field.

Amazon AI Review Summarizer: Launched on August 14, 2023, this feature on Amazon's website represents a significant stride in using AI for review summarization. This tool can be found at the top of the Amazon review section of every product accompanied by an indication that the content is "AI-generated from the text of customer reviews." The introduction of this feature has sparked discussions regarding its reliability and potential biases. As Amazon benefits financially from product sales, there is a concern that the AI's summarization process could be skewed to favor more positive interpretations, potentially misleading consumers. (The Verge, 2023).

FakeSpot: This free Chrome extension is designed to enhance transparency by identifying and filtering out fake reviews on e-commerce websites. Despite its noble intentions, FakeSpot's methodology exhibits significant flaws that can impact its utility and reliability. The extension tends to filter out many positive reviews, which, while often genuine, are treated with suspicion because products are more susceptible to fake positive reviews than negative ones. This bias results in disproportionately negative summaries, which can distort the perceived quality of a product. Such outcomes demonstrate the challenges in creating algorithms that accurately distinguish between genuine and manipulated content without introducing new biases. (Mozilla, 2024)

Reflecting on the existing tools such as Amazon's AI Review Summarizer and FakeSpot highlights the necessity for unbiased and accurate summarization technologies in e-commerce. These platforms demonstrate both the potential and the pitfalls of current review summarization efforts—where commercial biases and methodological flaws can skew the perceived value of products.

In contrast, our research employs publicly available large language models (LLMs) that are not specifically trained on Amazon review data and are free from commercial biases inherent in platform-developed tools. By leveraging these external LLMs, our study aims to critically evaluate the accuracy of Amazon's AI-generated summaries and to establish a benchmark for transparency and impartiality in review summarization. This approach not only challenges the effectiveness of existing commercial summarization tools but also contributes to the broader discourse on the reliability and ethics of AI applications in consumer markets.

2 Approach

For our research we looked specifically at the Home and Kitchen section of Amazon using a publicly available Amazon dataset (Yupeng Hou, 2023). We sampled 200 random products and extracted the product title, product description, overall star rating and the list of the 50 most recent product reviews. We also scraped amazon's website to obtain the product's AI generated review summary that Amazon integrated in the review section.

2.1 Models

Bart-large-cnn: The bart-large-cnn model is an adaptation of the BART (Bidirectional and Auto-Regressive Transformers) model which was released by Facebook back in 2019, which itself is a blend of bidirectional and autoregressive transformers. This particular variant has been fine-tuned using CNN-based architectures, specifically designed to enhance its capabilities in text summarization tasks. The model leverages a diverse set of text data, typically sourced from large datasets commonly used in NLP for summarization tasks, like CNN/Daily Mail for news article summarization. This model is relatively large, consisting of numerous transformer layers that enable it to capture complex textual relations and generate coherent summaries. It was trained on extensive datasets, often comprising millions of documents, to effectively learn summarization (Lewis et al., 2019).

flan-t5: flan-t5 stands for "Fine-tuned Language Net using T5 architecture," which extends the capabilities of the original T5 model, released by Google in 2020. T5, or "Text-to-Text Transfer Transformer," was designed to treat every NLP task as a text-to-text problem. flan-t5 advances this approach through instruction tuning to enhance performance on specific tasks like summarization. This model is trained on a vast array of text sources, employing datasets that are both broad and domain-specific, ensuring versatility across various NLP tasks. flan-t5 is an exceptionally large model, boasting millions to billions of parameters, allowing it to manage and synthesize extensive and diverse data effectively. The extensive training corpus includes a wide spectrum of internet text, often scaling to terabytes of data, which equips flan-t5 with the ability to generate precise and contextually accurate summaries (Chung et al., 2022).

Product Category	Review Samples	bart-large-cnn	flan-t5
Positive Product	"This pot is the go to pot for almost everything. It cooks evenly and is easy to clean. I will be buying other sizes." "Heavy duty. Does not feel like cheap junk. Not as easy to clean as teflon but it also won't make you sick." "Well made, looks great, perfect size to cook small meals." "Very sturdy and well made. It works so much better than the single ply pot I replaced."	The pot looks great and the lid has a nice mirror shine. Cuisinart is a known brand with a lifetime warranty that you know will be around to honor it. The pot makes perfect gravy, it is perfect for cleaning, and the size is great for small meals. For $30, I feel I got about 95% of the utility for like 20%.	Great quality, great price. I'm a fan of the 3-ply pans.Great for reheating food on the stove.
Negative Product	"Simple awful. The pot coating peeled after several uses. I used Teflon or wooden utensils so shouldn't have happened. Did not last any time. Very dissatisfied" "I really like this pot." "This is an example of absolute garbage. I cleaned the pot upon receipt and before I could even use, it has rust everywhere." "Love the size for small meals." "Well it has already started rusting on the basket handle. It boils over making a mess, kept bailing water out of it but kept still boiling over. I am bummed"	The strainer DOES NOT maintain a center of gravity - it keeps tilting to one side - (ergo) pasta falls under the strainer. The metal pieces that hold the straining basket are rusted, this pan is not worth the price! i used this pot maybe 3 times.... The outside handle is cracked and the drainer is starting to rust...i paid $40.00 for this pot and only used it like 3 times	Not so great the basket tips and is hard to get ahold of

Table 1: Comparison of Review samples and their bart-large-cnn review and flan-t5 review.

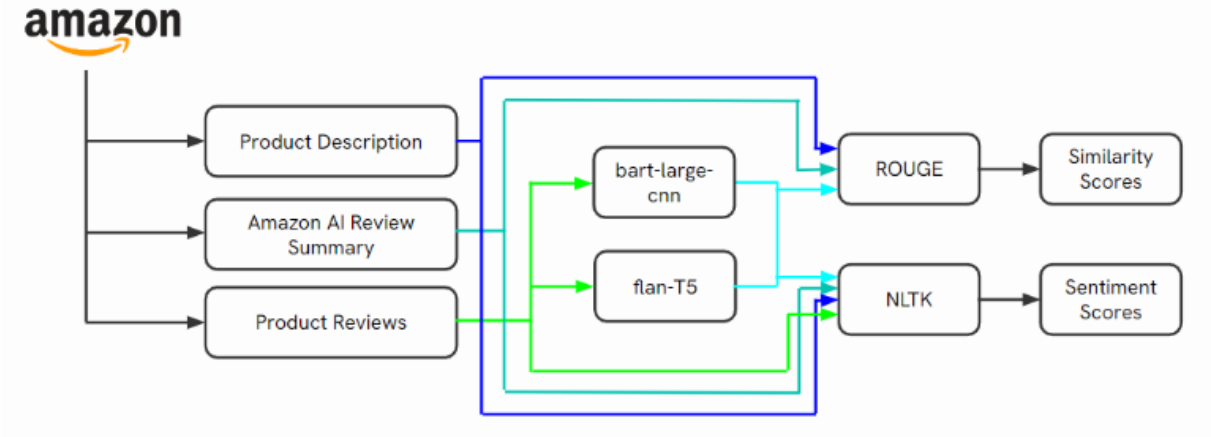

Figure 1: Pipeline

2.2 Summary Results

We looked at our results by categorizing the dataset we extracted into two different categories.

Product Category	Star Rating
Positively Reviewed Product	4.0 - 5.0
Negatively Reviewed Product	< 4.0

Table 2: Dataset Categorization and Labeling

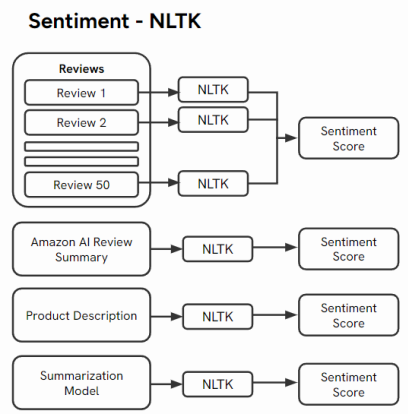

Table 3: NLTK Pipeline

3 Evaluation

To analyze the results of the model summarizations, we focus our evaluation based on two metrics: word similarity and sentiment. The word similarity provides a more concrete guideline of how similar each component is to each other. However, we also assess the sentiments in order to get a value that assesses the reviews in a more abstract manner. The sentiment of the product description, Amazon AI generated review summary, bart-large-cnn generated review summary, flan-t5 generated review summary and the average sentiment of the extracted reviews provides an assessment of the correlation in overall sentiments between the different sources of information. These contribute to the assessment of how faithful and transparent Amazon's product descriptions and AI generated reviews are.

We examine the sentiments of each component using NLTK's sentiment analysis score, measuring the overall sentiment of the reviews, the description and the summaries. Text similarity metrics like ROUGE-1 is then used to measure the unigram word similarity between the bart-large-cnn summaries and the product descriptions, the bart-large-cnn summaries and the Amazon AI generated summaries, the flan-t5 summaries and the product descriptions, and the flan-t5 summaries and Amazon AI generated summaries.

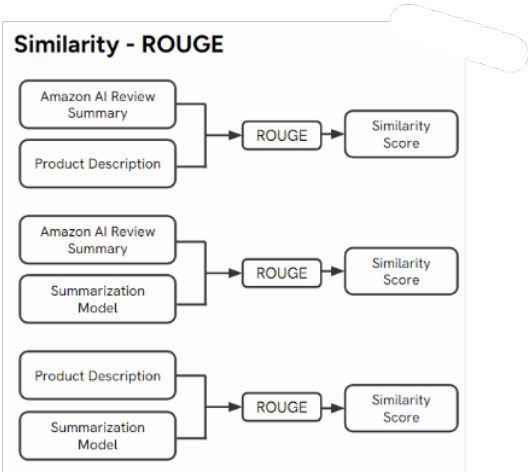

Table 4: ROUGE Pipeline

3.1 Result Analysis

After using NLTK's sentiment analysis and ROUGE to evaluate the different components of our research, we were able to get more insight on the summaries generated with bart-large-cnn and flan-t5. Looking at the summaries created by bart-large-cnn, when compared to both the product description and the Amazon

Category Negative Product	Positive Product
Amazon AI Summary 0.9403	0.9607
Product Description 0.7717	0.7506
Unprocessed Reviews 0.1774	0.9206
bart-large-cnn -0.52	0.9753
flan-t5 -0.1508	0.8885

Table 5: Sentiment score results

Category Negative Product	Positive Product
Bart-large-cnn vs Amazon AI Summary 0.2362	0.2735
Bart-large-cnn vs Product Description 0.2875	0.1403
Flan-t5 vs Amazon AI Summary 0.1449	0.1333
Flan-t5 vs Product Description 0.098	0.0833

Table 6: ROUGE-1 similarity score results

AI summary, the ROUGE-1 scores were much higher than the scores when compared to the summaries created by flan-t5 by 50-100%. The greater unigram word overlap can be attributed to the on average longer summaries that bart-large-cnn generates compared to the flan-t5 summaries, naturally incorporating more key words in the generated summaries. Bart-large-cnn is a model geared more towards understanding text in a greater context and specifically trained to generate text and perform well with text summarization tasks therefore producing summaries with unigram overlap scores of 0.14 - 0.3 whereas flan-t5 produced summaries with unigram overlap scores of 0.08 - 0.14. Flan-t5 was better at capturing the overall sentiment of the product reviews without any instruction but based on its large size and versatility, the model could output a more elaborate summary with further instruction and fine-tuning. The sentiment scores otherwise matched our hypothesis at the beginning of our research, with the product descriptions on average having higher scores, usually greater than 0.7, regardless of the actual sentiment of the product reflected in the reviews. The Amazon AI generated review also had higher scores compared to the summaries generated by bart-large-cnn and flan-t5, suggesting a slight positive bias. Overall our findings reveal a lack of transparency within the product descriptions, a slight positive bias in Amazon's AI generated reviews and support the use of models like bart-large-cnn and flan-t5 to help improve efficiency and transparency in the online marketplace.

4 Limitation and Qualitative Analysis

Considering the size of the Amazon marketplace and the model's input limitations, our research only sampled from a small section of products and reviews focused on not only one specific category but the most recent time period. We are also wary of the fact that our project focused on using the most recent 50 reviews for each product, something that could cause bias in the results if these were all part of mass generated fake reviews, or if the quality of the product during this time period had any irregular shifts. The use of these models to summarize thousands of reviews with a wide array of opinions also leads to the question of how it chooses only certain individuals' experiences to mention in the final generated summary and how it knows which ones are actually relevant. Our analysis revealed that higher ROUGE scores between the generated summaries of reviews and the product descriptions were primarily observed when negative reviews were involved. This is because negative reviews tend to be more detailed, often elaborating on specific aspects of a product. Such detailed discussions increase the overlap in content between the reviews and the product descriptions, which in turn boosts the ROUGE scores. This pattern was particularly evident when comparing summaries generated by bart-large-cnn and flan-t5 models with the original product descriptions. Despite the positive sentiment generally found in product descriptions and the contrasting negative sentiment in the reviews, the increased detail in negative feedback contributed to a higher un-

igram similarity. This suggests that the content-rich nature of negative reviews enhances their verbal alignment with product descriptions, leading to unexpectedly high ROUGE-1 scores. This often involves mentioning specific aspects of a product that are also highlighted in its description. This result suggests that models like bart-large-cnn and flan-t5 may have unintentionally generated summaries that are more negative as the negative reviews are often more elaborate and descriptive.

4.1 Next Steps

To enhance the scope and depth of our research, we outline several key advancements for the subsequent phase: Diversify the Models: We aim to expand our analytical framework by incorporating a broader range of models. This expansion will include not only different architectures but also more advanced models capable of handling larger datasets. Utilize Advanced Large-Scale Models: We plan to employ larger-scale models such as LLAMA, which are better suited for summarizing extensive collections of reviews. This approach will allow us to use a larger number of reviews in our summarization tasks. Broaden the Data Source: To mitigate bias introduced by focusing solely on a single marketplace, we intend to extend our dataset to include reviews from various sections of Amazon and other digital marketplaces. This diversification will provide a more balanced view of consumer sentiment across different platforms. Widen the Range of Summarization Tasks: Our evaluation of how these models perform on summarization tasks would expand between reviews summarization. Their performance would be tested on a variety of summarization tasks to assess their versatility and efficiency in different contexts. Enhance Evaluation Metrics: To systematically assess model effectiveness, we will develop a more robust evaluation framework. This framework will include advanced metrics that go beyond ROUGE and NLTK scores, aiming to capture relevance, coherence, and factual accuracy of the summaries. By implementing these steps, we aim to significantly advance our understanding of natural language processing applications and enrich the quality of automated summarization in practical settings.

5 Related Work

Study One: Benchmarking LLMs on the Semantic Overlap Summarization Task: This study explores how large language models (LLMs) perform in summarizing overlapping information from multiple documents, specifically focusing on the Semantic Overlap Summarization (SOS) task. The research utilizes the TELeR taxonomy to design diverse prompts to test the models' ability to capture the common themes in alternative narratives. The methodology and findings of this study provide insights into how LLMs handle complex, summarization tasks, which is analogous to summarizing diverse consumer reviews on platforms like Amazon (Salvador et al., 2024).

Study Two: LLMs in e-commerce: A Comparative Analysis of GPT and LLaMA Models in Product Review Evaluation: This study conducts a comparative analysis of two advanced LLMs, GPT-3.5 and LLaMA-2, to assess their effectiveness in understanding and analyzing sentiment within e-commerce product reviews. By evaluating these models on sentiment analysis tasks and comparing their effectiveness before and after specific fine-tuning, the study sheds light on the potential of LLMs to enhance customer satisfaction insights in e-commerce. This study showcases the practical applications of LLMs in interpreting complex consumer feedback and sentiment(Roumeliotis et al., 2024)

Study Three: Beyond Fake or Genuine – The Effect of Large Language Models (LLMs) on the Content and Sentiment of Product Reviews: This research investigates the impact of LLMs like ChatGPT on the creation and quality of product reviews. It hypothesizes that LLMs influence the content and sentiment distribution of reviews, promoting a focus on reviews with moderate ratings and richer informational content. This study's approach to evaluating the impact of LLMs on the generation and perception of product reviews and examines the broader implications of LLM-generated content in e-commerce environments and the impact on transparency and bias mitigation in consumer reviews.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Mozilla. 2024. How to use fakespot. https://blog.mozilla.org/en/products/how-to-use-fakespot/.

Konstantinos Roumeliotis, Nikolaos Tselikas, and Dimitrios Nasiopoulos. 2024. Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

John Salvador, Naman Bansal, Mousumi Akter, Souvika Sarkar, Anupam Das, and Shubhra Kanti Karmaker. 2024. Benchmarking llms on the semantic overlap summarization task. *Preprint*, arXiv:2402.17008.

The Verge. 2023. Amazon is now using ai to generate review summaries. https://www.theverge.com/2023/8/14/23831391/amazon-review-summaries-generative-ai.

Zhankui He An Yan Xiusi Chen Julian McAuley Yupeng Hou, Jiacheng Li. 2023. Bridging language and items for retrieval and recommendation.