User Experience Insights in LLM Platforms: A Comparative Analysis of ChatGPT and Claude

Xinyu Lin and Yujing Jiang

2024-12-10

GitHub: https://github.com/Colin0817/727-final-project.git

1 Introduction

1.1 Background

In recent years, the rapid advancement of Large Language Models (LLMs) has marked a significant breakthrough in the field of natural language processing. With their exceptional capabilities in text generation and comprehension, these models have found applications across a wide range of industries.

As the technology evolves, paid subscription services such as ChatGPT Plus and Claude Pro have emerged as a prominent business model for LLM platforms. These services not only enhance functionality but also attract a loyal user base by offering superior performance, exclusive support, and unique features tailored to user needs.

However, with increasing market competition, understanding users' experiences, emotional feedback, and core expectations from paid services has become critical. Gaining insights into these aspects is essential for optimizing service design, strengthening user engagement, and crafting effective market strategies.

1.2 Project Goal

This study seeks to analyze the user experience and emotional feedback of two prominent LLM paid services: ChatGPT Plus and Claude Pro. The goal is to explore differences in user satisfaction and functional requirements, with a focus on four key dimensions: **Usability**, **Price**, **Support**, and **Performance**.

By examining these dimensions, the study aims to identify the strengths and weaknesses of ChatGPT Plus and Claude Pro based on real user experiences. The findings will provide valuable data to support service enhancements and inform strategies for market competition.

1.3 Project Focus

In terms of user experience of paid LLM services, the following key questions are raised in this study:

1. **What is the user experience regarding LLM paid services? How do users emotionally perceive and respond to these services?**
2. **Which specific features and aspects of premium services are most valued by users?**

By answering these questions, the study will sort out the core features and feedback patterns of user experience and provide a scientific basis for optimizing paid LLM services.

2 Data Collection and Dataset Characteristics

This section describes our data collection from four major platforms: App Store, Google Play, Reddit, and YouTube. The dataset comprises user feedback and discussions about the paid subscription plans of ChatGPT

and Claude.

2.1 Platform Selection

We selected four platforms for their distinct characteristics in user engagement.

App Store is Apple's official application marketplace designed for iOS devices and stands as one of the world's earliest mobile application stores. Users can download both free and paid applications, provide ratings, and write reviews that help other users understand app quality and applicability. The platform's review system provides direct user feedback about app experiences.

Google Play, developed as Google's official application marketplace for Android devices, functions similarly to the App Store. Users can contribute reviews and ratings through their Google accounts, facilitating community interaction. Both these application marketplaces enable us to directly access user experiences with the ChatGPT and Claude apps.

Reddit is a community-driven social news and discussion platform that organizes content through "subreddits," covering topics from technology to entertainment. Users can post content and comments, with visibility determined by upvoting and downvoting mechanisms. Reddit's highly active communities facilitate in-depth discussions and experience sharing. Notably, Reddit's greater openness compared to App Store and Goodle Play makes it particularly suitable for collecting public opinions and analyzing various topics. Many users have shared comprehensive experiences and discussions about ChatGPT and Claude's paid plans on Reddit, making it a valuable data source.

YouTube, acquired by Google in 2006, is the world's largest video-sharing platform. It allows users to upload, watch, comment on, and share videos. YouTube's powerful search engine and recommendation algorithms help users discover relevant content. Users can like videos, comment, or subscribe to channels to interact with content creators. Videos related to LLM usage methods, experience demonstrations, updates, and development trends maintain considerable popularity on YouTube. Content creators often produce more diverse or in-depth videos to increase views and likes, resulting in potentially different characteristics from the other three platforms.

2.2 Data Collection Methodology

2.2.1 Data Collection

For systematic data collection, we implemented an API-based approach. The collection period spans from October 1, 2023, to October 1, 2024. To ensure relevance to our research objectives, we employed specific keyword filters for each platform.

For the application marketplaces (App Store and Google Play), we targeted reviews containing keywords such as *"plus," "chatgpt plus," "gpt plus," "upgrade," "premium," "chatgptplus," "gptplus,"* and *"gpt4"* for ChatGPT-related content.

For Claude-related content, we searched for terms including *"plus," "pro," "claude pro," "upgrade," "premium,"* and *"claudepro."*

For content-sharing platforms (Reddit and YouTube), we modified our keyword strategy to accommodate their broader discussion format. ChatGPT-related searches included *"gpt plus," "chatgpt premium," "chatgpt plus," "chatgpt upgrade," "gpt-4," "gpt4,"* and *"gptplus."* Claude-related searches focused on *"claude pro," "claude premium,"* and *"claude upgrade."*

2.2.2 API Implementation

Each platform required a specific API approach.

The App Store data collection utilized RSS API, based on XML or JSON standards. This public API allowed access to user reviews, though with certain extraction limitations.

For Google Play, we employed the Google Play Scraper Library, a semi-public tool requiring adherence to platform usage policies and anti-scraping guidelines.

Reddit's official API provided access to posts and comments through OAuth authentication, ensuring data collection complied with platform standards.

YouTube's official API required registration keys and operated under specific usage quotas, allowing access to video content and associated comments.

2.3 Dataset Characteristics

The final dataset comprises 11,148 comments distributed across platforms with the following distribution:

App Store yielded 193 ChatGPT-related reviews and 219 Claude-related reviews, providing insights from iOS users.

Google Play contributed a larger sample with 1,902 ChatGPT-related reviews and 1,109 Claude-related reviews from Android users.

Reddit, offering the largest sample, provided 3,559 ChatGPT-related comments and 2,006 Claude-related comments, reflecting its active community discussion environment.

YouTube maintained a balanced sample with 1,080 comments each for both ChatGPT and Claude-related content. The overview of collected data is presented in the following table:

Table 1 Overview of Data Collection Across Platforms

Platforms	Data Source	Data Quantity		Number of variables
		ChatGPT	Claude	
YouTube	YouTube API	1080	1080	4
Reddit	Reddit API	3559	2006	6
App store	RSS API	193	219	3
Google play	Google Play Scraper library	1902	1109	11

The dataset includes various data fields such as user ratings, textual comments, publication time stamps, and platform-specific metrics, enabling multifaceted analysis of user experiences and preferences regarding LLM model subscriptions.

3 Research Methods

3.1 Supervised Learning Method for Keyword Matching

In the study of text analysis and cause extraction, we employ a supervised learning approach based on keyword matching to identify user comments that address specific research dimensions. A multidimensional dictionary of key concepts wasemployed, encompassing four primary dimensions: Price, Performance, Support, and Usability. Below is a detailed description of each dimension:

Price: The Price dimension examines users' perceptions and evaluations of service pricing. Keywords were selected to encompass common price-related expressions found in user reviews, ranging from positive terms (e.g., "affordable," "worth," "inexpensive") to negative feedback (e.g., "expensive," "costly"). This approach captures various aspects of price perception, including its reasonableness, cost-effectiveness, and subscription fees. The keyword "subscription" is particularly focused on subscription-based payment models to analyze user feedback regarding long-term payment experiences.

Keywords: price, cost, expensive, cheap, value, affordable, worth, free, subscription, costly, inexpensive, budget.

Performance: The Performance dimension addresses the technical performance of a service, including system responsiveness, accuracy, reliability, and processing power. Keywords like "speed" and "lag" reflect concerns about service smoothness, while "accuracy" and "correctness" highlight users' expectations of precision and professionalism in service outputs. Additionally, terms such as "limit" and "delay" capture user experiences related to technical bottlenecks or service limitations.

Keywords: performance, speed, efficient, quality, accuracy, reliable, lag, fast, smooth, slow, output, limit, limited, delay, quick, correct, reliability.

Support: The Support dimension focuses on users' interactions with the service provider, emphasizing the efficiency of problem resolution, quality of technical support, and responsiveness of customer service. Keywords like "customer service" and "response time" reflect users' expectations regarding timely assistance, while terms such as "bugs" and "troubleshooting" indicate concerns about technical problem management.

Keywords: support, customer service, help, responsive, team, issues, bug, contact, solve, response time, feedback, documentation, helpful, troubleshooting, assistance, FAQ.

Usability: The Usability dimension highlights the ease and intuitiveness of user interactions with the platform. Keywords such as "easy" and "intuitive" reflect users' evaluations of user-friendliness, while "interface" and "design" focus on the aesthetics and functionality of the platform's design. Additionally, terms like "learning curve" and "customization" provide insights into users' experiences with adapting to and personalizing the platform.

Keywords: easy, intuitive, interface, design, user-friendly, simple, complex, accessible, learning curve, interaction, navigation, setup, customization.

We applied these keywords sequentially to various datasets from different application platforms (e.g., ChatGPT and Claude user reviews on the Apple App Store and Google Play Store) and social media platforms (e.g., user reviews of ChatGPT and Claude on YouTube and Reddit). Using string matching and pattern recognition techniques, we filtered and categorized review texts by dimension. The extraction process generated eight distinct comment sets for each platform and model, corresponding to the four research dimensions outlined above.

3.2 Sentiment Analysis Using the AFINN Lexicon

The AFINN lexicon, developed by Finnish researcher Finn Årup Nielsen, is a widely used sentiment analysis tool specifically designed for text sentiment evaluation. This lexicon comprises a collection of common English words and phrases, each assigned an integer sentiment score that reflects its emotional intensity and polarity.

To analyze sentiment, each word in the user comments is matched against the entries in the AFINN lexicon. The corresponding sentiment scores are then assigned to the comments based on the matched words. Using these scores, the comments are categorized into three sentiment states:

- **Positive**: Sentiment score > 0
- **Negative**: Sentiment score < 0
- **Neutral**: Sentiment score $= 0$

3.3 Twitter-RoBERTa Model Based on Hugging Face

To analyze social media data, we utilized the Twitter-RoBERTa model from CardiffNLP, available on the Hugging Face platform. This model is specifically optimized for social media text and supports fine-grained sentiment classification into three categories: **Positive**, **Negative**, and **Neutral**.

The Twitter-RoBERTa model is trained on a large corpus of tweet data, making it well-suited for sentiment analysis of user comments. After preprocessing the data, we applied the model to extract both the sentiment tendencies (Positive, Negative, Neutral) and their associated confidence scores.

Since the model outputs sentiment predictions using coded labels in the format `LABEL_x` (where x∈[1,2,3]), we mapped these labels to their corresponding sentiment tendencies for clarity.

4 Data Analysis

4.1 Analysis of Overall Sentiment and Reasons Among Users

This study integrates data from multiple platforms to evaluate the emotional attitudes of paid users toward ChatGPT and Claude. The analysis begins with an overall assessment of user sentiment and proceeds to examine the two models across four evaluation dimensions. Furthermore, the relationship between users' emotional attitudes and specific reasons is explored to understand the distribution of sentiments across the four dimensions.

4.1.1 What Is the Overall Emotional Attitude of Paid Subscribers Towards Both Models?

Figure 1 presents a comparison of the emotional attitudes of paid users toward ChatGPT and Claude. The data indicates that ChatGPT has a significantly higher proportion of positive sentiment compared to Claude, suggesting that most users express greater satisfaction with its service experience.

However, the difference in negative sentiment between the two models is minimal. ChatGPT exhibits a slightly higher proportion of negative sentiment, indicating that while most users are satisfied, some express greater dissatisfaction with its service compared to Claude.

Additionally, ChatGPT shows a marginally higher proportion of neutral sentiment. This finding may reflect that some users perceive its service as meeting, but not exceeding, their expectations.

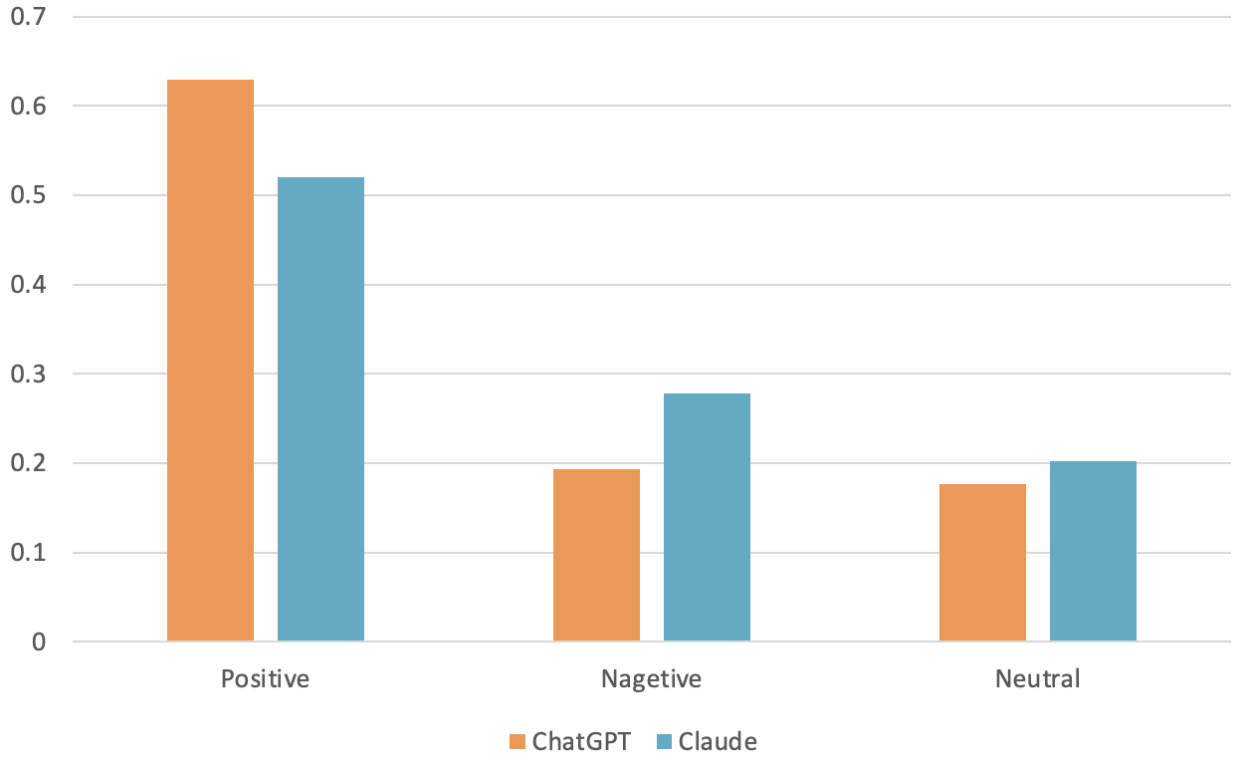

Figure 1 Sentiment Attitudes of Paid Users toward ChatGPT and Claude

4.1.2 What Are the Most Common Keywords Mentioned by Users of the Two Models?

Pricing is the most frequently mentioned factor by users, with ChatGPT receiving particularly high attention in this area (Figure 2). This indicates that users are more sensitive to pricing when evaluating ChatGPT and

have higher expectations for cost-effective services. In contrast, Claude shows a relatively low percentage of mentions regarding pricing, which may suggest that its users are either more accepting of its current pricing model or less concerned about pricing as a critical factor.

Regarding performance, ChatGPT is mentioned significantly more often than Claude. This trend reflects users' higher expectations for ChatGPT's responsiveness, accuracy, and stability.

Mentions of support services are also notably higher for ChatGPT compared to Claude. This trend could point to two potential interpretations: either users rely more heavily on ChatGPT's customer support services, or they provide more feedback on the platform's support quality, including suggestions for further improvement.

In the dimension of usability, mentions for Claude are slightly higher than for ChatGPT. This indicates that paid users are more focused on Claude's interface design, ease of operation, and overall user-friendliness. While ChatGPT users appear to provide less feedback on usability, this does not necessarily imply lower satisfaction.

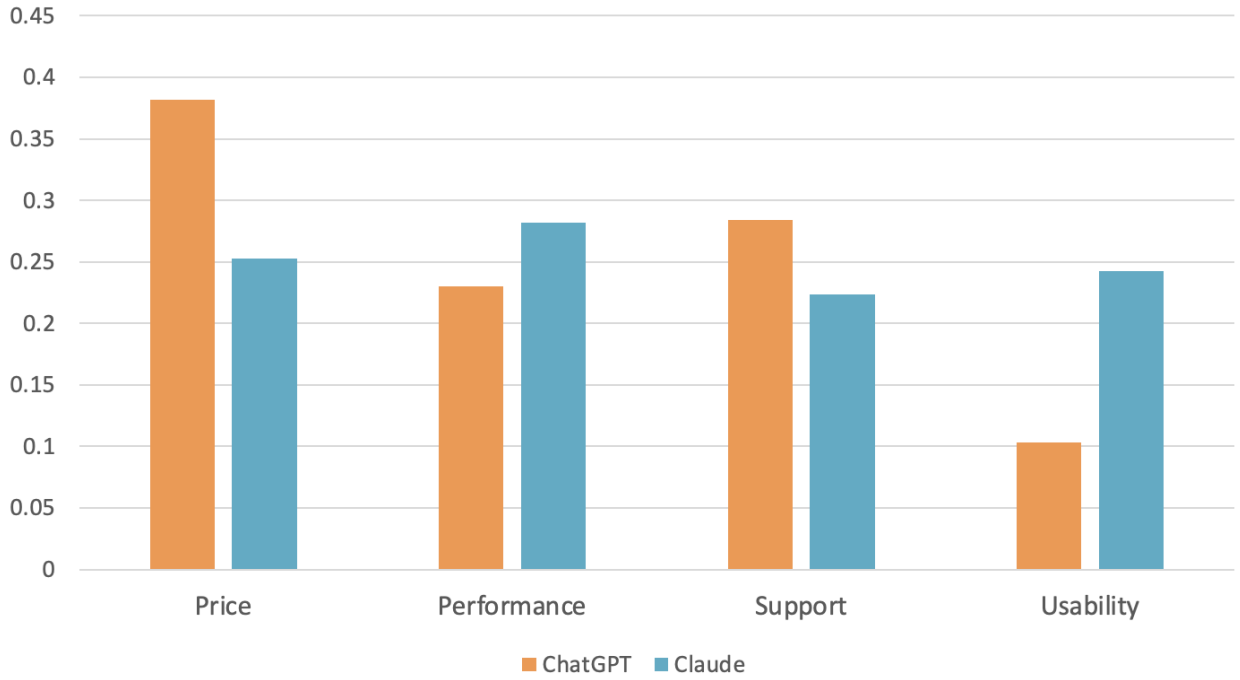

Figure 2 Proportions of Four Dimensions in Evaluations of ChatGPT and Claude

4.1.3 How Do Users Perceive the Dimensions of the Two Models?

To provide a clearer representation of the sentiment distribution among paid users across the four evaluation dimensions, this study extracted sentiment data for each dimension and created sentiment distribution charts for both ChatGPT and Claude (Figures 3 and 4).

1. **User Attitudes Toward Dimensions of ChatGPT**

Overall, despite some variation in sentiment, paid users generally expressed a positive attitude across all dimensions towards ChatGPT. A detailed analysis is as follows:

The Price dimension displayed a higher proportion of positive sentiment, indicating that many users believe ChatGPT offers good value for money and meets their payment expectations. However, Price remains the primary source of negative sentiment, with some users expressing dissatisfaction due to perceived high subscription fees or unmet expectations from value-added services. This suggests that pricing strategy continues to be a critical area for improvement.

Performance is the most positively recognized dimension, with significantly higher positive sentiment compared to the other dimensions. This reflects users' strong approval of ChatGPT's technical capabilities, particularly in terms of responsiveness, accuracy, and smoothness.

Sentiment in the Support dimension is largely positive, with users widely appreciating the responsiveness, problem-solving efficiency, and user experience optimization provided by customer service.

While the Usability dimension garnered fewer mentions, it still achieved a high proportion of positive sentiment. Users regard ChatGPT's interface design as intuitive and easy to operate, contributing to an overall user-friendly experience. Feedback suggests strong satisfaction with the platform's interface and functionality, even though this dimension received relatively less attention compared to others.

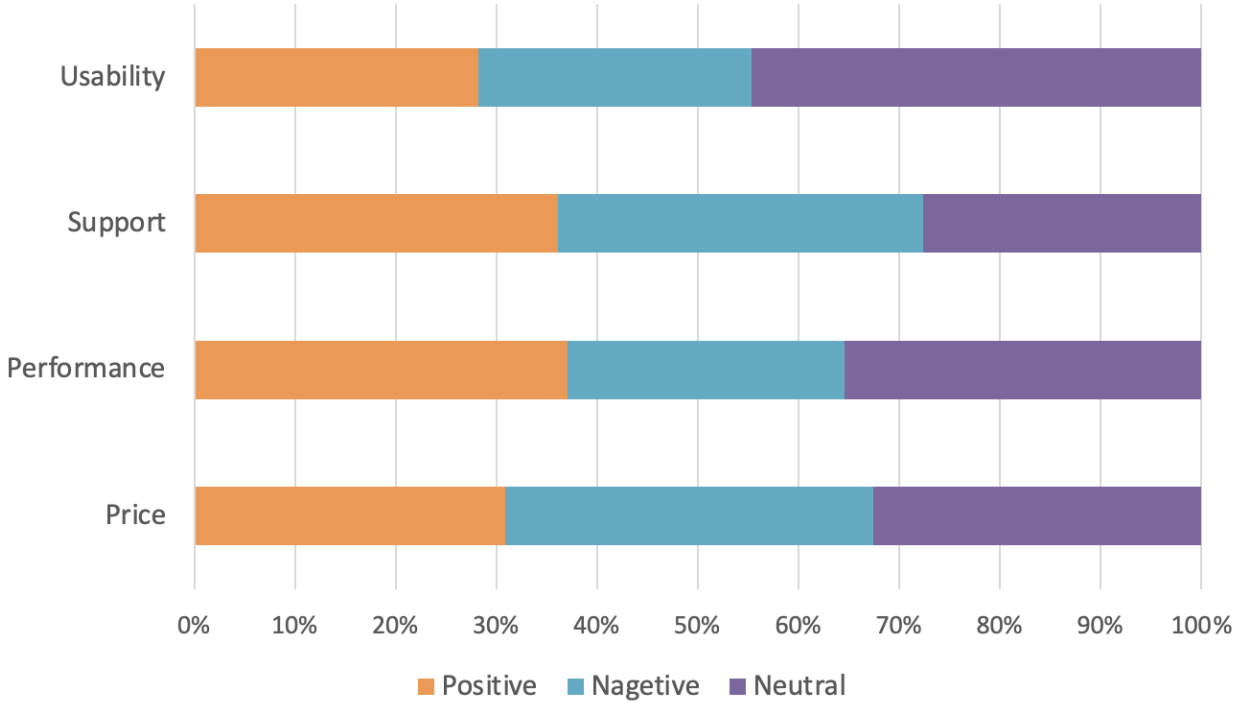

Figure 3 Sentiment Distribution Across Four Dimensions for ChatGPT Paid Users

2. **User Attitudes Toward Dimensions of Claude**

Claude has a similar proportion of positive sentiment in the Price dimension compared to ChatGPT. Additionally, the proportion of negative sentiment in this dimension is notably high, making pricing the primary source of dissatisfaction among paid users.

Similar to ChatGPT, the Performance dimension of positive attitudes accounts for 35% recognized by Claude users, reflecting its strengths in technical performance. However, Claude exhibits a relatively high percentage of negative sentiment in this dimension, suggesting persistent issues with stability or responsiveness that affect user satisfaction.

The Support dimension of Claude receives the highest recognition from paid users among all four dimensions. Claude has the lowest proportion of positive sentiment in the Usability dimension, accompanied by a notable level of neutral sentiment. This indicates that users find Claude's interface design and ease of operation to be average. Feedback suggests there is considerable potential for improving the platform's user-friendliness and ease of use.

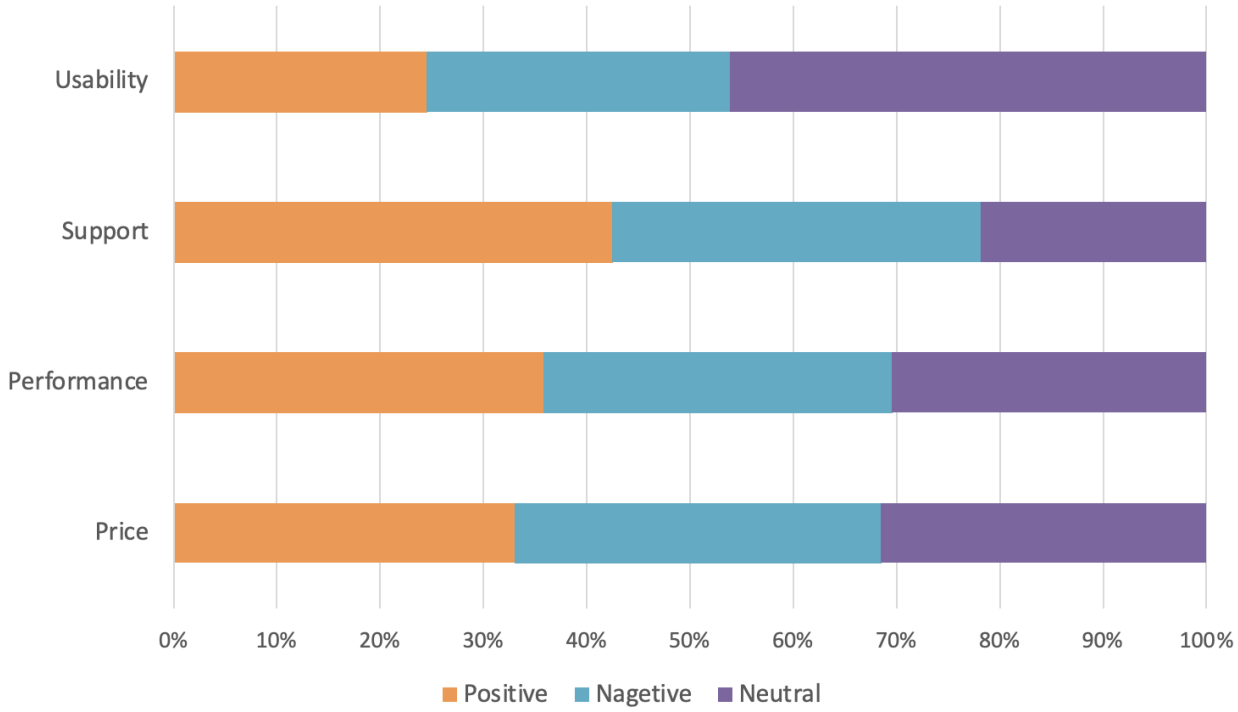

Figure 4 Sentiment Distribution Across Four Dimensions for Claude Paid Users

3. **Comparison of User Attitudes Across Dimensions**

A comparative analysis between ChatGPT and Claude reveals notable differences in user sentiment across the four dimensions. ChatGPT exhibits a similar proportion of positive sentiment compared to Claude, suggesting that users perceive it as cost-effective. In contrast, Claude has the highest proportion of negative sentiment in this dimension, highlighting its pricing strategy as a significant pain point in the user experience. Both models receive a high proportion of positive sentiment in the Performance dimension, indicating general satisfaction with their technical performance. However, Claude shows a higher proportion of negative sentiment.

Claude significantly outperforms ChatGPT in the Support dimension, with users expressing greater satisfaction with ChatGPT's customer support experience. Claude receives lower ratings from users in the Usability dimension, with feedback indicating that its interface design and interaction experience require further improvement.

4.2 Sentiment Comparison of ChatGPT and Claude Users Across Platforms in Four Dimensions

4.2.1 Do Users on Different Platforms Share Similar Attitudes Toward the Two Models?

Both ChatGPT and Claude receive high positive ratings across various platforms. The proportion of positive reviews on the App Store and Google Play is significantly higher than on Reddit and YouTube. This disparity may reflect differences in user review behaviors across platforms: users on the App Store and Google Play tend to leave reviews when they are either extremely satisfied or dissatisfied, often skewing toward extremes. In contrast, social media platforms like Reddit and YouTube feature more diverse and balanced reviews.

As shown in Figure 5, reviews on the App Store and Google Play are predominantly positive, with lower proportions of negative and neutral reviews. This suggests that users on these platforms are generally more satisfied with their overall experience of both ChatGPT and Claude. Conversely, on Reddit and YouTube, paid users show higher proportions of negative and neutral sentiment, with Reddit particularly standing out for its high percentage of negative reviews.

When comparing ChatGPT and Claude, we can find that ChatGPT has a higher percentage of positive sentiment than Claude on Reddit and YouTube. Claude, on the other hand, experiences a notably higher

percentage of negative sentiment, with the peak observed on Reddit.

On App Store and Google Play platforms, Claude slightly outperforms ChatGPT in terms of positive sentiment, indicating that its user experience on these platforms is relatively well-received. However, Claude still shows a higher proportion of negative sentiment compared to ChatGPT, suggesting that it may not fully meet user expectations in certain areas of detail or functionality.

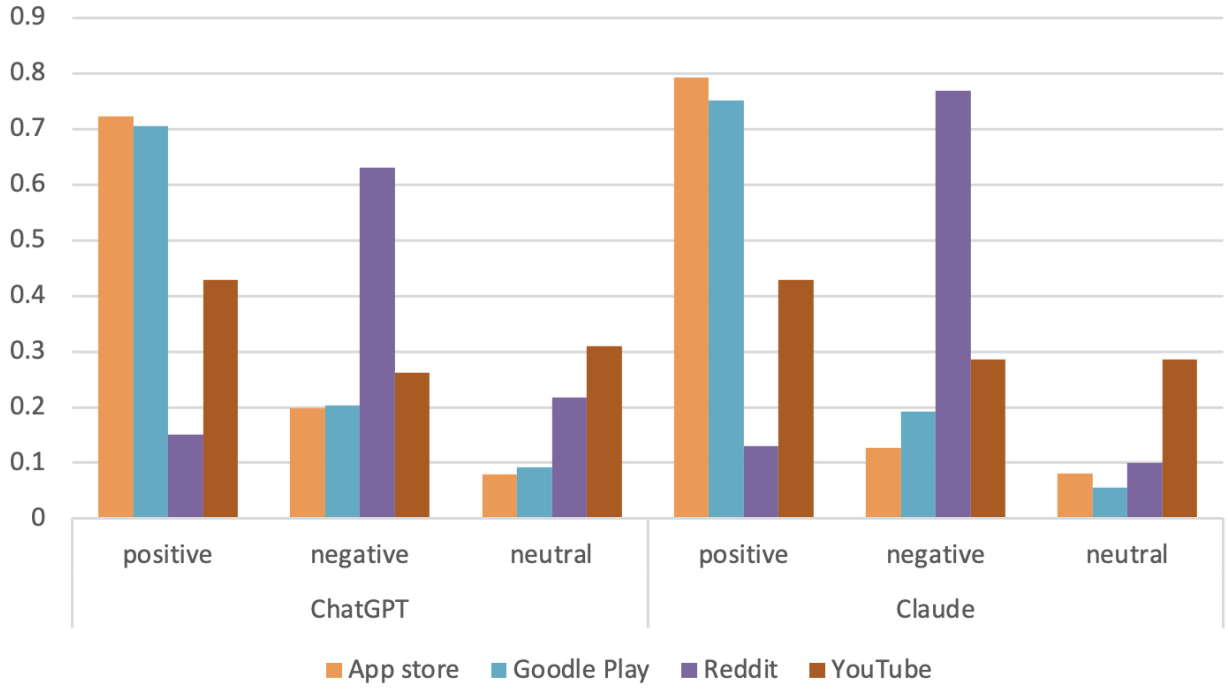

Figure 5 Sentiment Comparison of ChatGPT and Claude Paid Users Across Platforms

4.2.2 Do Users on Different Platforms Focus on Similar Dimensions?

Claude's mention rate is generally higher than ChatGPT's, particularly on the Google Play and Reddit platforms. This suggests that users are more focused on these three dimensions when evaluating Claude(Figure 6).

The mention rate for the Price dimension varies significantly across platforms, with Claude consistently receiving a much higher mention rate than ChatGPT. This indicates that pricing is the most critical factor for Claude's users, especially on Google Play and Reddit. Similarly, Claude outpaces ChatGPT in mention rates for Performance and Support dimensions, reinforcing the idea that users are focusing on these aspects more closely when reviewing Claude. Both ChatGPT and Claude exhibit lower mention rates in the Usability dimension, implying that users pay relatively limited attention to interface design and ease of operation.

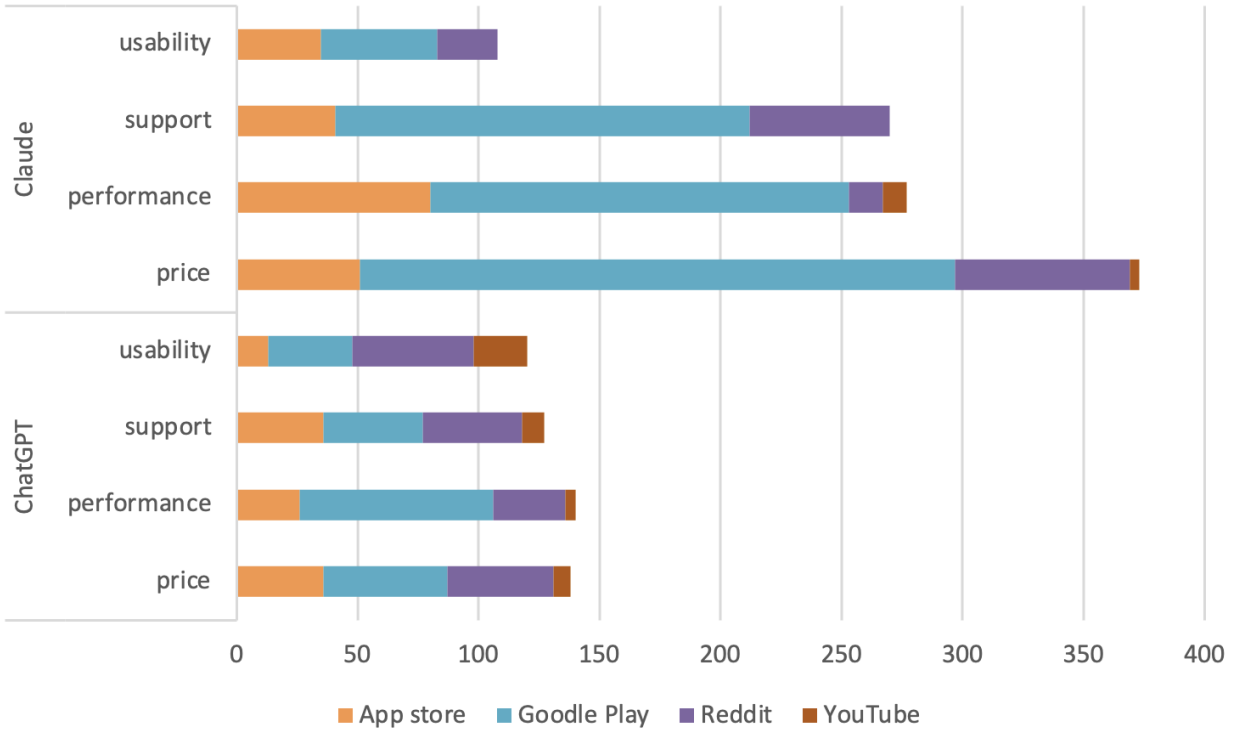

Figure 6 Cross-Platform Sentiment Distribution in Four Dimensions: Claude vs. ChatGPT

4.2.3 Do Users Across Platforms Share Similar Emotional Attitudes Toward the Dimensions of the Two Models?

To thoroughly understand the experiences of paid users and their reasons for choosing different LLM models, we analyzed the percentage distribution of sentiments (positive, neutral, and negative) across four dimensions and visualized the results using heat maps. These heat maps, based on user comment data from four platforms, provide insights into sentiment trends.

Figures 7 and 8 illustrate the distribution of positive sentiment across platforms and dimensions.

On the App Store and Google Play platforms, the proportion of positive user comments for the four dimensions of both ChatGPT and Claude is relatively high, as indicated by the green shading on the heat map. This reflects a generally high level of user satisfaction with the two models on these platforms. In contrast, on Reddit and YouTube, the proportion of positive sentiment is lower, represented by purple on the heat map. This suggests that users on these social media platforms are inclined to share less favorable views of both ChatGPT and Claude.

For ChatGPT, sentiment data from the App Store and Google Play platforms differ significantly from those of Reddit and YouTube, reflecting large fluctuations in user emotional feedback across platforms. In comparison, Claude's heat map shows a more even color distribution, indicating greater consistency in user sentiment across platforms.

Both ChatGPT and Claude demonstrate higher proportions of positive sentiment in the Usability dimension. Notably, ChatGPT outperforms Claude in positive Usability feedback across all platforms, reflecting stronger user satisfaction with its interface design and ease of operation. ChatGPT exhibits some platform-based differences in positive sentiment for the Performance dimension, with weaker performance specifically noted on the YouTube platform. Claude has a lower proportion of positive sentiment in the Price dimension compared to ChatGPT, reinforcing earlier findings that pricing remains a significant pain point for Claude users. Both models show similar positive sentiment proportions in the Support dimension, indicating a perceived balance in user support services between ChatGPT and Claude.

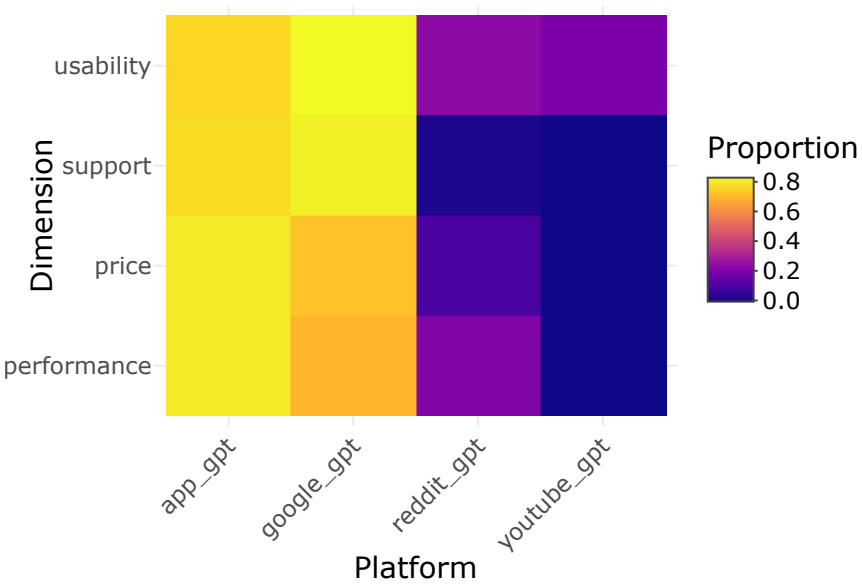

Figure 7 Positive Sentiment Proportions for ChatGPT

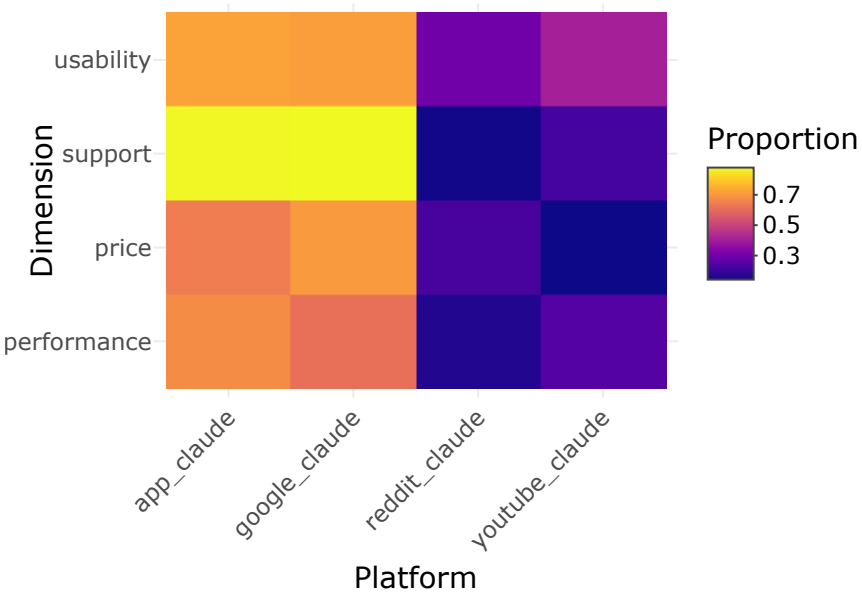

Figure 8 Positive Sentiment Proportions for Claude

Using the share of negative sentiment expressed by users towards ChatGPT and Claude across the four dimensions, heat maps (Figures 9 and 10) were generated to visualize the distribution of users' negative feedback across platforms.

The share of negative sentiment for ChatGPT varies significantly across dimensions and platforms, with some reaching nearly 1.0. On the YouTube platform, the Price and Support dimensions show the highest percentages of negative sentiment, suggesting that users are especially dissatisfied with these aspects.

Claude exhibits a lower share of negative sentiment compared to ChatGPT, indicating relatively higher user satisfaction. However, certain areas still require attention. On the YouTube platform, the Price dimension displays a high proportion of negative sentiment, nearing 0.6. This suggests that pricing remains a significant pain point for Claude users, who may feel that the cost of the service does not align with its perceived value.

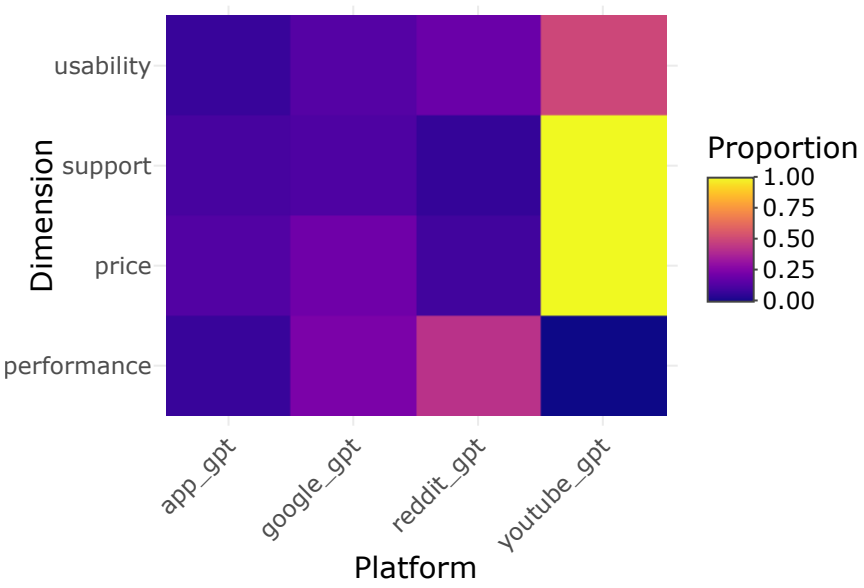

Figure 9 Negative Sentiment Proportions for ChatGPT

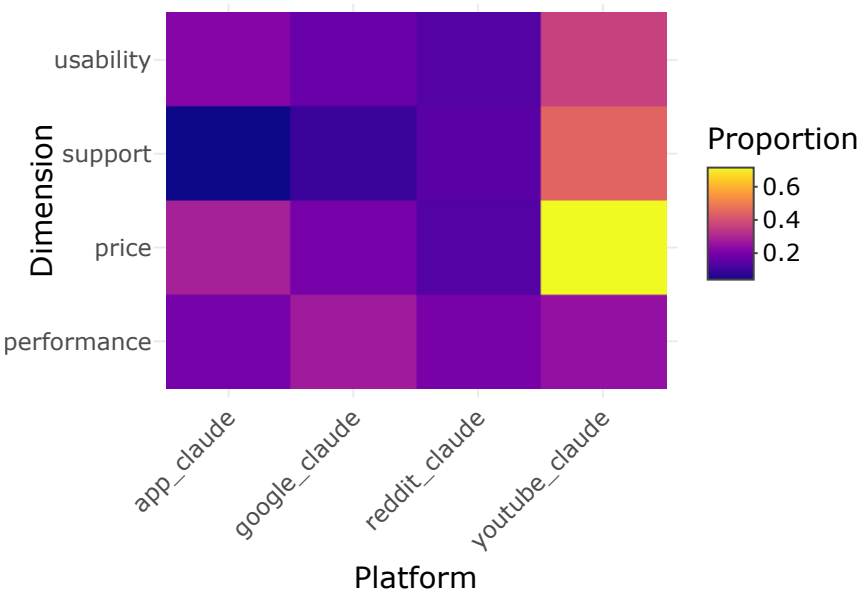

Figure 10 Negative Sentiment Proportions for Claude

5 Conclusion

5.1 Comparative Analysis of Core Features

The analysis of user comments across platforms reveals distinct patterns in user perceptions of ChatGPT and Claude, particularly regarding their core features and functionality. This section presents key findings, organized by major themes and platform-specific insights.

Overall, ChatGPT receives more positive feedback compared to Claude. The most significant positive comments about ChatGPT are focused on the performance and support dimensions, indicating that users are generally satisfied with its technical performance and the quality of assistance provided. On the other hand, Claude receives consistent positive feedback on its support features. User comments frequently highlight that Claude has unique advantages in customer support and problem solving.

A notable observation is that price is one of the most important factors for LLM users. Despite both models having the same monthly paid plans, users' perceptions of the price for ChatGPT and Claude differ significantly. The analysis shows that price is a major pain point for Claude, with many users expressing dissatisfaction. In contrast, ChatGPT enjoys a more favorable perception of its pricing, although it remains a concern for some users. This suggests that, for most users, ChatGPT's paid plans are seen as offering better value for money. This may be attributed to ChatGPT's strong performance in terms of output speed, quality, and reliability, which better meets the needs of LLM users, leading to an overall more positive paid experience.

5.2 Platform-Specific Comment Characteristics and Comparison

The analysis of user comments across four platforms—App Store, Google Play, Reddit, and YouTube—highlights distinct differences in user behavior, expectations, and feedback styles unique to each platform.

First, there is a clear contrast in the emotional tone of comments between application marketplaces and content-sharing platforms. On App Store and Google Play, users tend to provide either highly positive or negative feedback, with little emphasis on neutral or constructive criticism. This can be attributed to the review-based structure of these platforms, where users are focused on offering a general evaluation of the service. In contrast, Reddit and YouTube offer a wider range of emotional tones, with users providing more balanced feedback, including constructive criticism and recognition of both strengths and weaknesses. This diversity of sentiment makes feedback on Reddit and YouTube more valuable for identifying areas of improvement.

Another key difference lies in the focus on service dimensions. App Store and Google Play users primarily concentrate on performance and price, with the emphasis on price being especially strong on Google Play. In contrast, Reddit and YouTube users demonstrate a more balanced focus across all four dimensions, but place particular emphasis on support and usability for Claude. This indicates that Reddit and YouTube users are more critical of Claude's usability and customer support and are more likely to provide detailed feedback on these aspects, whereas users on App Store and Google Play focus more on the overall value and functionality of the service.4.3

5.3 Recommendations for Enhancing LLM User Experience

Based on the analysis, we can draw several key recommendations for enhancing user experience with LLM models.

First, it is essential to clearly identify and address the core needs of users. When developing paid plans, a strong focus should be placed on optimizing model performance and actively promoting these improvements. A reasonable balance between performance and pricing is critical, as users expect high performance and fair pricing. If users purchase a paid plan but are dissatisfied with the service, they are more likely to leave negative feedback, which can negatively affect the model's reputation.

Second, user support and interface design remain key areas for improving the overall user experience. Faster, more efficient support services can significantly enhance user satisfaction, while a more aesthetically pleasing, user-friendly interface makes the model easier to use. Offering both effective support and a well-designed, intuitive interface will contribute to positive user experiences and improve retention.

Finally, when collecting user feedback, special attention should be given to the insights shared on content-sharing platforms. Unlike application marketplaces, which often incentivize reviews and lead to either very brief comments or an inflated number of positive reviews, content-sharing platforms provide a more open environment where users can freely express a variety of opinions. These platforms tend to offer more balanced and objective feedback, which can help better identify pain points and areas for improvement that may be overlooked on review-based platforms.

5.4 Possible Limitations

While the analysis provides valuable insights into user perceptions of ChatGPT and Claude, there are several limitations that should be considered:

5.4.1 Difficulty in Determining Paid Users

One of the main challenges in this analysis is the difficulty in identifying whether users are paying customers. Many platforms do not provide clear distinctions between free and paid users in the comments. Without this information, it becomes challenging to isolate feedback from users who are paying for the service, potentially skewing the understanding of user satisfaction across different segments.

5.4.2 Limited App Store Reviews

Due to restrictions on scraping reviews from the App Store, only a limited number of reviews were accessible for analysis. This restriction means that the analysis of App Store comments is based on a smaller sample size, which could affect the reliability of the insights drawn from this platform compared to others with more extensive feedback.

5.4.3 Comparison Limited to Two LLM Models

Another limitation is that the analysis only compares two LLM models, ChatGPT and Claude. While this provides useful insights into the user experience with these two models, it does not allow for a broader comparison across other available LLMs. Including additional models in the analysis could help identify trends that apply to the broader LLM market and provide more comprehensive recommendations for improvements.

For future research, it would be valuable to collaborate with companies and use their API to distinguish between paid and free users, providing more detailed insights into each group's experience. Additionally, to overcome the App Store data limitations, web scraping or other data collection methods could be used to gather a larger sample of reviews. Lastly, the analysis could be expanded to include more LLM models, such as Gemini or other LLM models, to offer a more comprehensive comparison across the LLM landscape.