



Krzysztof Krej

mgr inż., Wydział Ekonomiczno-Informatyczny,
Wyższa Szkoła Ekonomii i Informatyki
w Krakowie (WSEI),
email: kkrej@wsei.edu.pl
ORCID: 0009-0005-0880-5457

ANALYSIS AND UTILIZATION OF LARGE LANGUAGE MODELS FOR DRAWING INSIGHTS FROM VIDEO GAME REVIEWS

ANALIZA I WYKORZYSTANIE DUŻYCH MODELI JĘZYKOWYCH DO
WYCIĄGANIA WNIOSKÓW Z RECENZJI GRY KOMPUTEROWEJ

Key words: artificial intelligence, natural language processing, game development, GPT models, review analysis

Słowa kluczowe: sztuczna inteligencja, przetwarzanie języka naturalnego, rozwój gier, modele GPT, analiza recenzji

JEL Classification: O33, D83

Abstract

This study investigates the use of GPT models for analyzing video game reviews. The research develops a methodology for efficiently gathering, classifying, and interpreting player feedback using GPT models and explores strategies for integrating insights into game development processes. Experiments using player reviews for "House Flipper 2" demonstrate the effectiveness of GPT models in extracting meaningful insights. Implications for game development, challenges associated with large-scale automated analysis, and future research directions are discussed.

Streszczenie

Praca bada wykorzystanie modeli GPT do analizy recenzji gier wideo. Badanie opracowuje metodologię efektywnego gromadzenia, klasyfikowania i interpretowania opinii graczy z wykorzystaniem modeli GPT oraz bada strategie integracji uzyskanych spostrzeżeń z procesami rozwoju gier. Eksperymenty z wykorzystaniem recenzji graczy gry "House Flipper 2" pokazują skuteczność modeli GPT w wydobywaniu znaczących spostrzeżeń. Omówiono implikacje dla rozwoju gier, wyzwania związane z automatyczną analizą na dużą skalę oraz przyszłe kierunki badań.

INTRODUCTION

In the rapidly evolving landscape of the video game industry, understanding and effectively utilizing player feedback has become a critical aspect of game development. With the rise of digital distribution platforms and social media, players now have unprecedented opportunities to share their opinions, experiences, and suggestions through various channels, such as user reviews, forums, and social media posts. This wealth of user-generated content presents both opportunities and challenges for game developers seeking to improve their products and meet player expectations.

Traditionally, analyzing player feedback has been a time-consuming and labor-intensive process, often relying on manual methods such as reading and categorizing individual reviews. However, the sheer volume of available data has made this approach increasingly impractical, necessitating the development of more efficient and scalable solutions. In recent years, the emergence of advanced natural language processing (NLP) techniques and large language models (LLMs) has opened up new possibilities for automated analysis of player feedback [5].

This paper explores the potential of using LLMs, specifically GPT models, for the analysis and interpretation of video game reviews. By leveraging the capabilities of these powerful language models, this study demonstrates how game developers can gain valuable insights from player feedback, enabling data-driven decision-making and ultimately leading to improved game quality and player satisfaction.

The primary objectives of this research are threefold:

- To evaluate the effectiveness of GPT models in classifying and analyzing video game reviews, comparing their performance to traditional manual analysis methods.

- To develop a methodology for utilizing GPT models to efficiently gather, classify, and interpret player feedback, identifying best practices and techniques for optimizing the analysis process.
- To explore strategies for integrating insights derived from GPT-based analysis into the decision-making processes and development cycles of video games, with the goal of enhancing product quality and player satisfaction.

To achieve these objectives, the study involved a series of experiments using a dataset of player reviews for the game "House Flipper 2." By applying various data filtering, contextual querying, and iterative refinement techniques, this study demonstrates how GPT models can effectively extract meaningful insights from player feedback, identifying key trends, issues, and suggestions for improvement.

Furthermore, I discuss the implications of the findings of this study for game development, highlighting the potential benefits of data-driven decision-making, targeted community engagement, and continuous improvement based on player feedback. I also address the challenges associated with large-scale automated analysis, such as data quality, contextual understanding, and ethical considerations.

Finally, future directions are outlined for research and development in this field, including the refinement of LLMs for enhanced accuracy, the expansion of multilingual analysis capabilities, and the incorporation of user engagement data. By pursuing these avenues, I believe that game developers can further harness the power of LLMs to gain valuable insights from player feedback, ultimately leading to the creation of more engaging, satisfying, and successful video games.

THEORETICAL BACKGROUND

In the era of digitalization and the growing significance of digital media, video games have become an essential element of popular culture, engaging millions of users worldwide. This mass reception makes games not only a form of entertainment but also an important object of scientific research. Ludology, as a field dedicated to the study of games and play, is expanding rapidly, shedding light on various aspects of this media form – from game narratives and mechanics to their impact on society and culture.

In the context of the constantly evolving video game market, with particular emphasis on the dynamic environment of production and the

reactions of the player community, there is a need to utilize modern technologies for data analysis and interpretation. Such analysis is particularly important in times when game production increasingly relies on the principles of data-driven development, where design decisions and product development directions are based on analyzed data.

The Role of User Feedback in Game Development

User feedback significantly influences the life cycle of a game [1]. It can be used to infer what updates players expect, determine whether it is a good time to release additional paid or free content for the game (DLC), and what that content should include. With a thriving game, the development team needs a constantly updated task list, the implementation of which will translate into real value for players. The use of data as a basis for decision-making in game development (data-driven development) has become a key aspect of decision-making in game production and development.

Utilization of Reviews and Opinions in the Decision-Making Process

Player reviews published in game stores, such as Steam, are an invaluable source of information for game creators and publishers [9]. They not only provide evidence of user acceptance or rejection of a given production but also contain suggestions, highlight problems, and introduce community expectations and preferences into the discussion. Understanding this data and effectively analyzing it is key to success in an industry that, due to its dynamic nature, requires quick adaptation and a creative approach to development.

The Role of AI and NLP Technologies

The incredibly rapid development of artificial intelligence capable of processing natural language has led to the emergence of tools capable of advanced analysis of vast textual datasets. One of the most promising examples of these technologies is the GPT (Generative Pre-trained Transformer) models developed by OpenAI. They can play a crucial role in ana-

lyzing review texts, offering unique capabilities in identifying patterns that can influence the direction of game development.

Literature Review

Article [9] presents the results of a study proving that video game reviews are a rich source of content. It identifies nine themes that regularly appear in game reviews, but it does not address methods for automatically analyzing their content.

Utilizing LLMs (Large Language Models) for the sentiment analysis of video game reviews is a relatively new topic that the literature devoted to it is surprisingly modest.

Sentiment analysis itself is a relatively well-described topic. For example, [4] presents a method of aspect-based sentiment analysis (ABSA), which allows for the assessment of user attitudes towards previously released video games. The authors focused on a three-part game series, analyzing user reviews to identify the most frequently mentioned words, called "aspects". These aspects were then examined in terms of sentiment, which was found to be highly correlated with user ratings.

The described analysis is particularly valuable for game developers who want to understand which elements of their games were well-received and which met with negative reactions. The research methodology focuses on understanding which aspects of the game (such as plot, combat, or characters) are most important to users and how they are evaluated. By applying this knowledge, developers can better tailor their future projects to player expectations.

An example of the game series studied is "Dragon Age," which, despite sales successes, received mixed reviews from players. This study allowed for a deeper understanding of how design decisions influenced user perception of the game and which aspects were criticized or praised in individual parts of the series.

The conclusions from the analysis can serve as a tool for better understanding player expectations and for creating more engaging and satisfying games by focusing on the aspects most important to users. Furthermore, the study emphasizes the importance of directly incorporating user opinions into the game design process to avoid discrepancies between developer expectations and player experience.

Similarly, [8] and [1] discuss the use of natural language processing (NLP) techniques to draw conclusions. However, the techniques described

here, such as word counting or sentence length analysis, which can be used to study language complexity in game reviews, also do not provide sufficient information for making business decisions regarding further game development.

In [3], a highly systematic approach is employed, sequentially performing sentiment analysis, binary classification, multi-label text classification, and topic modeling. However, even such a complex approach can only quantify game playability based on aspects like gameplay, usability, and functionality, without offering specific recommendations to improve the game.

A similar approach was used in [2], where the authors even attempt to read the causal relationships between user arguments and purchase decisions. However, in this case as well, the raw lexical approach only allows for a simple classification of possible product issues and does not allow for the discovery of those that were not anticipated by the experiment authors.

Paper [5] is the only one I could find concerning the use of LLMs for review analysis. It illustrates the higher effectiveness of this approach. Worth noting is the example of a cynical review cited in the paper ("Great for uninstalling. 11/10 that I would uninstall again."). In the case of such reviews, lexical methods often incorrectly assess them as positive, and only the use of LLM allows for reading the true intention. Unfortunately, none of the discussed articles talks about an approach that allows drawing general, specific, previously unknown conclusions about which specific game elements (such as mechanics or a specific type of content) are most praised or criticized by players.

In the further part of the work, we will focus on the methodology of using GPT models in the context of games, analyzing both the potential and limitations of such an approach, and then present the research results and their implications for the future of video game development.

METHODOLOGY

Usage Scenarios

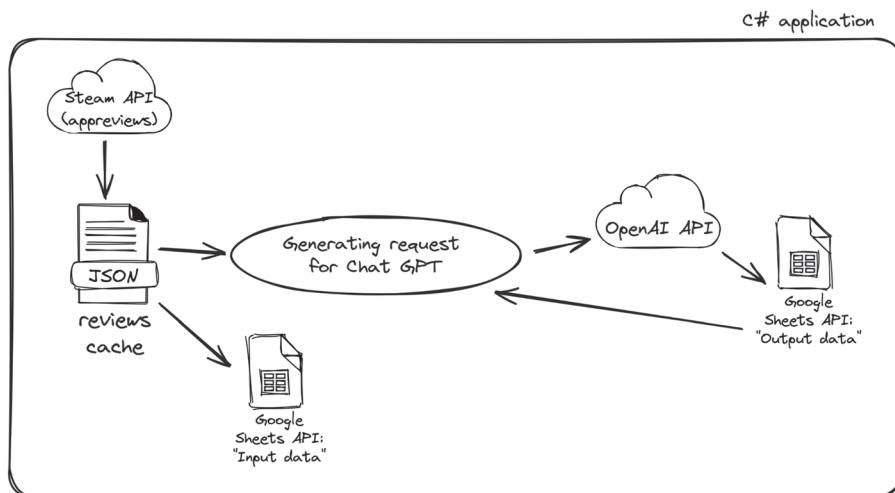
There are two principal approaches for utilizing the tools discussed. The first approach involves an unstructured analysis of Steam reviews. In this scenario, a database of reviews is sent to the ChatGPT with a request to summarize them. The model is tasked with analyzing positive and nega-

tive reviews separately, identifying recurring themes, and then producing a summary that highlights aspects of the game that are most praised, most criticized, and suggestions for changes (which appear in both positive and negative reviews). In this scenario, the reviews are not filtered.

The second approach entails searching for reviews related to a specific topic. Initially, the ChatGPT generates a list of key phrases related to the topic. Subsequently, the input data is filtered using these phrases. A new query containing the filtered database of reviews is then sent to the ChatGPT, the results of which are recorded in a spreadsheet.

Data flow

Figure 1. Flowchart of the game review analysis application



Source: Author's own elaboration.

In the described software system, multiple external and internal data sources were integrated using a C# application to handle and analyze user-generated content from Steam reviews. The process begins with the C# program querying the Steam API to fetch player reviews. Once retrieved, these reviews are temporarily stored locally in a JSON file, serving as a cache to optimize response times for subsequent requests. If the same data request is made again, the system checks this cache before hitting the Steam API, thus reducing unnecessary network calls and load on the

external API. The Steam API is rate-limited so it is possible to download about 10 reviews per second. In case of game I'm working on, House Flipper 2, currently there are more than 6000 reviews to download, so without cache it would take 10 minutes each time. Adding JSON cache solved this problem

Following local caching, the data is then exported to a Google Sheet labeled "Input Data", allowing users to conveniently access and review the raw data.

For analytical processing, the C# application generates request containing reviews data from Steam, and sends it to the OpenAI API (ChatGPT), where natural language processing is employed to analyze the content of the reviews. The results of this analysis are then recorded in a separate Google Sheet named "Output data" completing the data flow cycle. On user request these data can be used as input data for OpenAI API again, for example to summarize results.

This system not only provides a streamlined way to collect and analyze user reviews but also facilitates an efficient data management protocol by leveraging caching mechanisms and integrating with cloud-based spreadsheet tools for accessible data visualization and result storage.

Data filtering

Practical tests have demonstrated that when the GPT model is presented with thousands of comments and asked to extract details concerning a specific issue, it tends to include summaries of the overall content and other aspects not explicitly requested. Consequently, the initial step involves basic filtering. For instance, when the summary of negative comments suggested that some players experience motion sickness while playing the analyzed game, the following question was posed:

"In House Flipper 2, players may sometimes experience mild motion sickness or nausea. I would like to review player comments that address this issue. Could you provide me with a list of keywords and phrases that I could use to filter the reviews to find approximately 1% of them that discuss this issue?"

The response received was:

"motion sickness, nausea, dizziness, uncomfortable, queasy, eye strain, headache, discomfort, difficulty adjusting, camera movement, flicker effect, rapid scrolling, vertigo, vomit".

This selection was surprisingly detailed and apt, surpassing what I might have devised on my own.

After filtering reviews with these keywords, 43 reviews discussing this topic were identified.

These filtered reviews are then sent to the ChatGPT with an additional instruction to focus solely on the analyzed aspect, in this case:

"What do players complain about regarding motion sickness or dizziness in the game and how would they like the game to be changed to address these issues? I am not interested in opinions about other aspects of the game at this moment".

Only by combining keyword-level filtering with a clear query to the ChatGPT explicitly requesting the exclusion of other topics were satisfying results obtained.

It is important to note that in this case, the keyword filter restricted the analysis to reviews in English – this need not be the case. Keywords can be in multiple languages, and the ChatGPT handles most European languages well. Unfortunately, Asian languages often fall short, leading me to frequently exclude them from analysis as the reviews translated from these languages by artificial intelligence often do not make sense after translation. This is unfortunate, as Asian players constitute as much as 16% of the analyzed games' audience. Fortunately, many of them write reviews in English, allowing us to effectively analyze them.

Preparing query for OpenAI

The described C# application generates a query for the ChatGPT based on the text written by the user in the following manner:

First, a system message is sent with the content: "You are an expert answering questions about the computer game House Flipper 2 based on player opinions". This query works because ChatGPT knows what the game House Flipper is – it was easy to verify this in a conversation with it. For less known products or other games, it is necessary to provide a context description so that the AI knows the topic of the reviews and can properly understand them.

In the next message, a question is asked to the system. It is constructed according to the following pattern: "Write an answer to the question based on the provided player reviews. The question is:

{question}

Reviews to analyze:

{reviews}"

This approach ensures that the AI has the necessary context to understand the question and the reviews it needs to analyze. By providing the AI with a clear instruction to answer the question based on the given reviews, relevant and focused responses can be obtained.

It is important to note that for less popular games or products, additional context may need to be provided to the AI. This could include a brief description of the game or product, its main features, and any other relevant information that would help the AI better understand the topic of the reviews.

By following this query preparation pattern, the C# application can effectively communicate with the OpenAI API and obtain meaningful insights from the player reviews. This structured approach allows for the automation of the review analysis process, saving time and resources while providing valuable data for game developers to make informed decisions.

Utilizing the Appropriate Large Language Model Version

This study exclusively utilized the ChatGPT (Generative Pre-trained Transformer) provided by OpenAI. My choice was driven by the desire to employ the most intelligent tool available. Numerous tests, such as those found at <https://lifearchitect.ai/iq-testing-ai/>, clearly suggest that GPT is among the top performers. The only comparable competitor at this level is Claude 3, which, however, is not yet available in Poland, hence my decision to use GPT.

Most users are familiar with ChatGPT from its web version. While this version is an excellent tool for beginners, it is considerably limited compared to what the Application Programming Interface (API) allows. As of this writing in May 2024, the most significant models are GPT-4 and GPT-4 Turbo. Choosing between these two models presents an intriguing issue. OpenAI states that "With 128k context, fresher knowledge and the broadest set of capabilities, GPT-4 Turbo is more powerful than GPT-4 and offered at a lower price". Indeed, GPT-4 Turbo appears to be superior:

- It uses fresher training data from December 2023, while GPT-4 has a knowledge base up to September 2021. ([source](<https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4>)),
- It allows for the analysis of over ten times more data: 128,000 tokens (equivalent to about 300 pages of text), compared to 8,192 to-

kens for GPT-4. ([source](https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4)),

- It is more than twice as cost-effective, with a million tokens of input and output costing respectively \$10 and \$30 for GPT-4 Turbo, but as much as \$30 and \$60 for regular GPT-4. ([source](https://openai.com/pricing)),
- Additionally, GPT-4 Turbo is faster, deemed more accurate, has a larger vocabulary, and better handles adaptive learning. ([source](https://medium.com/@olivierlaurence81/gpt-4-vs-gpt-4-turbo-whats-the-difference-94f195bfc423)).

Given these advantages, does it make sense to use the standard GPT-4? It turns out that sometimes it does. Some data suggest that this model performs better in tasks that do not require working with large amounts of data or within a limited time frame.

For instance, it excels in generating code ([source](https://aider.chat/docs/benchmarks-1106.html)) and achieves higher scores in the SAT – the standardized American test used for college admissions in the United States ([source](https://twitter.com/wangzjeff/status/1721934560919994823)).

Consequently, the study employed GPT-4 Turbo for queries related to data analysis, but use the standard GPT-4 for generating keywords and summarizing results.

DATA ANALYSIS

Reviews summarising

The primary application of this approach was to automate writing summaries of reviews. The analyzed game receives hundreds to thousands of reviews each month, many of which span several pages of text. Reading them in their entirety could take about ten hours per month for anyone wanting to familiarize themselves with the content. Moreover, many reviews are very similar, so this task could become monotonous. ChatGPT excels at writing summaries. Below is a sample summary: *Summary of 500 reviews of House Flipper 2 (80% positive) collected from 2024-03-19 to 2024-05-02: Positive reviews:*

- *Overall, very good ratings for the game, especially for the graphics and the possibility of creative play.*

- *Players appreciate the new features and tools available in the game, such as furniture matching and coloring.*
- *Sandbox and tutorial rated positively for accessibility and freedom in creativity.*
- *Some users emphasize that the game is much better than the first part of the series, praising it for its improvements and new possibilities.*
- *The game is relaxing and provides many hours of enjoyable entertainment; great for long sessions.*
- *Users appreciate regular updates from developers adding new content and features.*
- *Praised for its graphics – especially for details and aesthetics.*

Negative reviews:

- *Many users report technical problems such as bugs, game crashes, and performance problems – especially on higher graphics settings.*
- *Users report a lack of decoration items; indicate that the game has less content compared to the first part, especially it is devoid of garden options, which were available in the DLC for the first part.*
- *There is frustration with the game feeling “unfinished” and concerns about whether future updates will be free or paid.*
- *Some players, compared to the first part of the series, believe that House Flipper 2 is a step backwards.*
- *Disappointed with the lack of some features or tools that were available in previous versions of the game.*
- *A few complaints also concern the user interface, which can be confusing or less intuitive than players expected.*
- *No support for Steam Workshop, which limits the possibilities of modding the game.*

Overall, House Flipper 2 is generally well received for its relaxing nature and creative elements, with regular updates adding value to the game. However, technical bugs, a limited number of available decorations compared to the previous entry in the series, and concerns about future content cloud the overall positive image of the game.

The conclusions from this summary align very well with what I have determined myself from reading the game reviews and discussions on forums. Although this might only be my subjective evaluation, another point provides more analytical confirmation of the utility of the discussed solution.

Comparison with Data Obtained by Traditional Methods

I use various tools, including *canny.io*, to collect feedback from players.

Canny.io is a SaaS (Software as a Service) platform designed to help companies gather, organize, and analyze customer feedback efficiently. It is primarily used by product teams to track user requests and feature suggestions, enabling a more user-centric approach to product development. Canny.io provides tools for creating feedback boards where users can submit their feedback, vote on existing ideas, and engage in discussions.

Thanks to this, there is a clear list of items that players would like me to add to the game, as well as mechanics and gameplay features they consider most important.

Interesting conclusions can be drawn by comparing the data collected from *canny.io* with what the discussed model can extract from player reviews.

Below is an excerpt from interaction with the discussed application:
What topic would you like to summarize the reviews on?

> What specific mechanics & gameplay features are players asking for? List specific ones, in bullet points from the most popular request.

I am analyzing 500 reviews posted between 2024-03-19 and 2024-05-02.

Players have been asking for various mechanics and features in House Flipper 2 to improve their gaming experience. Here is a list of the most popular requests by reviews:

- 1. **Greater variety of decorations and furniture:** – Many reviews note the lack of variety in available furniture, decorations and finishing options.*
- 2. **Improving tools and painting mechanics:** – Players are dissatisfied with the current wall painting tools, requiring additional upgrades to make painting faster and more efficient.*
- 3. **Game stability and optimization:** – Some players experience problems with game performance, including crashes and FPS drops, which especially applies to players with modern computers.*
- 4. **Before and After Slideshow:** – Users are asking for the return of the function of comparing the condition of the house before and after renovation, which was one of the favorite elements of the first part of the game.*
- 5. **Greater interaction and influence on auctions:** – Players would like to have a greater influence on the process of selling houses at auctions, so that they can better manage the expectations and choices of potential buyers.*

The table below compares responses from GPT based on game reviews and submissions created by players on canny.io:

Table 1. Comparison of GPT responses based on game reviews with player submissions on Canny.io

GPT Response	Equivalent on Canny.io
Greater variety of decorations and furniture	Separate category on canny.io with >10K votes
Improving tools and painting mechanics	Top #5 and top #15 requests on canny.io are about adding new tools and improving existing ones
Game stability and optimization	Canny.io is not used for collecting information about bugs and the need for optimization
Before and After Slideshow	Top #1 request from players on canny.io, with 2.2K upvotes
Greater interaction and influence on auctions	Top #4 request from players on canny.io, with 1.4K upvotes

Source: Author's own elaboration based on analysis of House Flipper 2 reviews and Canny.io data.

Out of the five most popular submissions on Canny.io, three are also described as the most frequently mentioned in game reviews. This is a very good result. Manual checking showed that the remaining two most popular submissions on canny.io, namely:

- "Undo" button for Story Mode (1.4K upvotes),
 - Show us how many rooms still have quests left in jobs (1.5K upvotes),
- were not mentioned in game reviews by players during this period. This makes sense, as while these are important features for players, their absence probably does not impact the joy of playing as much as the other mentioned items do.

Examples of Contextual Analysis

Beyond searching for recurring requests, comments, and suggestions in reviews, an important way of using the discussed tools is the possibility of conducting instant research on a given topic. The methodology employed this for analyzing necessary improvements to many mechanics and aspects of the game. I will cite an example where it was possible to successfully identify particularly useful conclusions:

List comma-separated key phrases you want to search for in reviews:

> buyer, buyers, auction, auctions

Found 106 reviews matching the key phrases. This represents 1% of all reviews.

What topic would you like to summarize the reviews on?

> What do players complain about in-game buyers and how would they like to see them changed? (By "buyers", I mean virtual characters that appear in the game and to whom the player can sell a renovated house. I am not interested in opinions about aspects of the game other than merchants at the moment)

Analysis of 106 reviews posted between December 15, 2023, and April 19, 2024:

Players in House Flipper 2 have expressed dissatisfaction with the in-game buyer interaction, specifically the auctioning process. Critically, the feedback highlights that buyer engagement is not as dynamic or interactive as it was in the first game. Buyers previously showed reactions and provided feedback that catered to specific preferences during house renovations, making the auction more engaging and strategic. Players miss this element as it added depth and personality to the buying process.

Players suggest that improving buyer interaction in House Flipper 2 could involve reintroducing more detailed buyer preferences and feedback during auctions. They would appreciate a system where buyers provide specific requests or express distinct likes and dislikes that influence the auction outcomes. This change would not only bring back the vibrant character interaction found in the first game but also enhance the strategic aspect of house flipping, as players would need to tailor their renovations to meet different buyer demands effectively.

End of analysis

In the discussed example, it was possible to identify specific, clearly described changes that need to be implemented for players to more appreciate the discussed system. Given how crucial this aspect of the game is, Extensive time was spent analyzing player requests, forum discussions, reviews, and votes on canny.io. It is worth noting that this 154-word summary of 106 reviews concerning the topic hits the mark and clearly suggests what our company should do to make the game better.

Prioritizing player requests

To evaluate the effectiveness of ChatGPT in prioritizing player requests, I utilized data from Canny.io, a platform that allows players to submit and vote on feature requests for the game. The data consisted of a list of player requests, ranked by the number of votes each request received.

The research involved several experiments where I provided ChatGPT with a subset of the most recent player reviews and asked it to rank the feature requests based on the information contained in these reviews. Then I compared ChatGPT's rankings with the actual rankings from Canny.io using Spearman's rank correlation coefficient.

Spearman's rank correlation coefficient is a non-parametric measure of the monotonicity of the relationship between two datasets. It assesses how well the relationship between two variables can be described using a monotonic function. The coefficient ranges from -1 to $+1$, where $+1$ indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

The study employed Spearman's rank correlation coefficient for several reasons:

1. It does not assume a linear relationship between variables, making it suitable for comparing rankings.
2. It is less sensitive to outliers compared to Pearson's correlation coefficient.
3. It can handle ties in rankings, which are common in user-generated data like feature requests.

In our experiments, I first calculated the Spearman's rank correlation coefficient between the actual rankings and an alphabetically sorted list of the feature requests. As expected, this yielded a coefficient close to zero (-0.031), indicating no meaningful correlation.

Then I conducted multiple trials where I provided ChatGPT with different subsets of the most recent player reviews. The Spearman's rank correlation coefficients between ChatGPT's rankings and the actual rankings ranged from 0.16 to 0.27 . This variability is expected due to the non-deterministic nature of ChatGPT.

I consider coefficients above 0.15 to be favorable, given several factors:

- The dynamic nature of the data, as player preferences may shift over time.
- The small differences in the number of votes between some feature requests, making their relative positions highly sensitive to minor changes.

These results suggest that ChatGPT can provide meaningful insights into player priorities based on an analysis of player reviews, as evidenced by the positive correlation between its rankings and the actual rankings from Canny.io.

Changing Trends

Finally, it's important to note that while analyzing as large a database of reviews as possible seems tempting, some of them become outdated over time. For example, during the first weeks after the game's release, players most frequently reported in reviews a lack of curtains. This was a big surprise for me, but given the overwhelming number of requests to add curtain models, I added them to the game within two months of the release. Since that day, the problem has disappeared, and reviews related to it have also vanished. Initially, when I analyzed all reviews every time, not just the current ones, the informational noise associated with repeating already solved problems was distracting. Ensuring that only the latest few hundred reviews are analyzed made the entire process faster, cheaper, and more accurate.

DISCUSSION

Comparison with Other Methods

Large Language Models (LLMs) have proven to be significantly more effective than traditional Natural Language Processing (NLP) methods in analyzing video game reviews [1, 2, 6]. They can extract more accurate insights from review texts, answer specific questions, limit the analysis to a given context, work with multiple languages, and effectively handle irony, which negative reviews are often filled with.

Compared to manual text analysis, using LLMs has numerous advantages. First and foremost, it is incomparably faster – analyzing 300 pages of reviews takes between 10 and 30 seconds, while it would take a human an average of 8-12 hours. Moreover, LLMs do not display emotions, do not succumb to fatigue or boredom, which is exceptionally important. For game developers, negative reviews are the most crucial, and reading negative reviews of their game can easily lead to mental breakdowns for an average developer. Reading a one-page summary of these comments prepared by an LLM is incomparably better in terms of emotional hygiene.

Strategies and Techniques for Optimized Analysis

To effectively utilize Large Language Models (LLMs) for analyzing video game reviews, several strategies and techniques can be employed to optimize the process and obtain the most relevant insights.

- **Data Filtering:** Before feeding the reviews to the LLM, it is crucial to filter the data based on specific keywords or phrases related to the topic of interest. This helps to narrow down the analysis to the most relevant reviews, reducing noise and improving the quality of the insights generated. Filtering can be done based on keywords related to specific game mechanics, features, or user experiences.
- **Contextual Querying:** When interacting with the LLM, it is essential to provide a clear and specific context for the analysis. This can be achieved by formulating queries that explicitly state the focus of the analysis and exclude irrelevant aspects. For example, instead of asking for a general summary of the reviews, the query should be tailored to address specific questions or concerns, such as "What do players think about the game's difficulty level, and how do they suggest improving it?"
- **Iterative Refinement:** The analysis process should be iterative, allowing for the refinement of queries based on the initial results. If the first round of analysis yields broad or incomplete insights, the queries can be reformulated to dive deeper into specific aspects or to clarify ambiguities. This iterative approach ensures that the final insights are comprehensive and address the most critical points.
- **Multilingual Analysis:** On one hand, automatic translations of Asian languages often pose significant challenges. On the other hand, it is common for reviews from players in specific regions to have lower ratings. This may be due to issues affecting only players in those regions, such as poor quality translations into a given language or (in the case of China) lack of access to certain internet resources caused by the government firewall. The optimal solution seems to be asking the same question twice, once for all comments and a second time for comments in languages with the highest translation quality. This may involve creating separate keyword lists for filtering and adapting the queries to suit the linguistic and cultural nuances of each language.
- **Temporal Analysis:** Analyzing reviews over time can reveal valuable insights into how player opinions and experiences evolve. By seg-

menting the reviews into different time periods (e.g., launch week, first month, post-update), developers can track changes in sentiment, identify emerging issues, and assess the impact of updates or patches on player satisfaction.

- **Integration with Other Data Sources:** To gain a comprehensive understanding of player feedback, the insights derived from LLM analysis should be integrated with data from other sources, such as customer feedback collecting platforms, social media discussions, and bug reports. This holistic approach provides a more complete picture of the game's reception and helps prioritize development efforts.

By employing these strategies and techniques, game developers can optimize their use of LLMs for analyzing player reviews, ensuring that the insights generated are accurate, relevant, and actionable. This enables data-driven decision-making in game development, ultimately leading to improved player satisfaction and game success.

Integrating LLM-Derived Insights into Decision-Making

Incorporating insights from Large Language Models (LLMs) into the decision-making process is a highly effective way of utilizing player feedback. By integrating LLM-derived insights, developers can make data-driven decisions that align with player preferences and expectations, leading to improved game quality and player satisfaction.

Insights derived from LLMs can help developers prioritize game updates and improvements based on the most pressing issues and desired features identified through player feedback. However, it is crucial to balance addressing player concerns with maintaining the game's original design vision. Players often express their dissatisfaction but may be unaware of its true cause. Additionally, players tend to resist change – for example, if the developers modify the user interface, players may express discontent even if the change is for the better, as the altered interface may not align with their established habits.

LLM-derived insights can also guide developers in crafting targeted communication and community engagement strategies [3]. After implementing changes based on these insights, assessing the impact of updates on player satisfaction is essential for continuous improvement.

Challenges in Large-Scale Automated Analysis

While Large Language Models (LLMs) offer numerous benefits for analyzing player feedback, developers must navigate several challenges when implementing large-scale automated analysis:

- **Data Quality and Noise:** Player feedback data can be noisy, containing irrelevant or ambiguous information, hindering accurate analysis.
- **Contextual Understanding:** LLMs may struggle to grasp the context and nuances of player feedback, particularly with game-specific terminology or references. Developers must provide sufficient context and validate insights to ensure accuracy.
- **Sarcasm and Irony Detection:** Players often employ sarcasm, irony, or humor in their feedback, which can be challenging for LLMs to accurately interpret.
- **Language Differences:** Translations may not always capture the precise meaning or sentiment of the original text, leading to potential misinterpretations.
- **Bias and Subjectivity:** Player feedback is inherently subjective, reflecting individual preferences and experiences.
- **Handling Contradictory Feedback:** Players may express conflicting opinions about the same aspect of the game. If something is mentioned only in negative reviews, it may be liked by many users but not considered important enough to mention in positive reviews.
- **Privacy and Data Security:** Collecting and analyzing player feedback raises privacy and data security concerns. Developers must adhere to data protection regulations and implement appropriate safeguards.

To mitigate these challenges, developers should combine automated analysis with human oversight and interpretation, regularly refine the analysis process, incorporate feedback from domain experts, and stay updated with the latest advancements in NLP techniques.

By proactively addressing these challenges and continuously refining their approach, developers can harness the power of LLMs to gain valuable insights from player feedback while ensuring data integrity, reliability, and ethical use.

CONCLUSION

Summary of Findings

The research demonstrates that Large Language Models (LLMs), particularly GPT models, can be effectively utilized for the analysis and interpretation of video game reviews. Key findings include:

- LLMs outperform traditional Natural Language Processing (NLP) methods in extracting accurate insights, answering specific questions, and handling contextual nuances in game reviews.
- Compared to manual text analysis, LLMs offer significant advantages in terms of speed, emotional neutrality, and the ability to process large volumes of data.
- Effective strategies for optimizing LLM-based analysis include data filtering, contextual querying, iterative refinement, multilingual analysis, temporal analysis, and integration with other data sources.
- LLM-derived insights can be successfully integrated into the decision-making process for game development, enabling data-driven prioritization of updates and improvements based on player feedback.
- Challenges in large-scale automated analysis include data quality and noise, contextual understanding, sarcasm detection, language differences, bias and subjectivity, handling contradictory feedback, and privacy and data security concerns.

Implications for Game Development

The findings of this research have significant implications for game development:

- Data-driven decision-making: LLM-based analysis of player reviews enables developers to make informed decisions based on actual player feedback, leading to improved game quality and player satisfaction.
- Prioritization of updates and improvements: Insights derived from LLMs can guide developers in prioritizing game updates and improvements, focusing on the most pressing issues and desired features identified by players.

- Enhanced community engagement: LLM-derived insights can inform targeted communication and community engagement strategies, fostering a stronger connection between developers and players.
- Balancing player feedback and design vision: While incorporating player feedback is crucial, developers must strike a balance between addressing player concerns and maintaining the game's original design vision.

FUTURE DIRECTIONS

Based on the research findings, several future directions can be explored:

- Refining the Model for Enhanced Accuracy: Continued research and development of LLMs specifically tailored for video game review analysis can further improve the accuracy and reliability of the insights generated.
- Expanding Multilingual Analysis Capabilities: Investing in the development of more robust multilingual analysis capabilities, particularly for Asian languages, can enable developers to gain insights from a broader range of player feedback.
- Incorporating User Engagement Data: Integrating LLM-based analysis with user engagement data, such as playtime, in-game behavior, and social media interactions, can provide a more comprehensive understanding of player experiences and preferences.
- Ethical and Privacy Considerations: As the use of LLMs in game development grows, it is crucial to address ethical and privacy concerns related to the collection, analysis, and use of player feedback data. Establishing clear guidelines and safeguards will be essential to ensure responsible and transparent practices.

By pursuing these future directions, game developers can further harness the potential of LLMs to gain valuable insights from player feedback, ultimately leading to the creation of more engaging, satisfying, and successful video games.

BIBLIOGRAPHY

1. Jeffrey, Robert, et al. *"The wisdom of the gaming crowd"*. *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction*

- in Play*. 2020. [2] Kunaefi, Anang, and Masayoshi Aritsugi. "Extracting arguments based on user decisions in app reviews". *IEEE Access* 9 (2021): 45078-45094.
2. Li, Xiaozhou, Zheyang Zhang, and Kostas Stefanidis. "A data-driven approach for video game playability analysis based on players' reviews". *Information* 12.3 (2021): 129.
 3. Strååt, Björn, Harko Verhagen, and Henrik Warpefelt. "Probing user opinions in an indirect way: an aspect based sentiment analysis of game reviews". *Proceedings of the 21st International Academic Mindtrek Conference*. 2017.
 4. Vigiato, Markos, and Cor-Paul Bezemer. "Leveraging the OPT Large Language Model for Sentiment Analysis of Game Reviews". *IEEE Transactions on Games* (2023).
 5. Zagal, Jose P., Noriko Tomuro, and Andriy Shepitsen. "Natural Language Processing for Games Studies Research". *Journal of Simulation & Gaming (S&G), Special Issue on Games Research Methods* 43.3 (2011): 353-370.
 6. Zagal, José P., Amanda Ladd, and Terris Johnson. "Characterizing and understanding game reviews". *Proceedings of the 4th international Conference on Foundations of Digital Games*. 2009.