Mining Player Experience Trends From Game Reviews Using Large Language Models

Supriya Dutta
Aalto University
Espoo, Finland
supriya.dutta@aalto.fi

Joel Oksanen
Aalto University
Espoo, Finland
joel.oksanen@aalto.fi

Jaakko Väkevä
Aalto University
Espoo, Finland
jaakko.vakeva@aalto.fi

Shamit Ahmed
Aalto University
Espoo, Finland
shamit.ahmed@aalto.fi

Markus Kirjonen
Aalto University
Espoo, Finland
markus.kirjonen@aalto.fi

Perttu Hämäläinen
Aalto University
Espoo, Finland
perttu.hamalainen@aalto.fi

Abstract

How have player experiences changed over the years? For instance, have there been general shifts in what kinds of emotions players experience and express? We probe these questions with help of recent methodological advances in psychology and Large Language Models (LLMs), in particular the possibility to predict Likert-scale responses based on free-form text. Applying this at scale to three player experience questionnaires (PXI, CORGIS, AESTHEMOS) and 152143 Metacritic user reviews from years 2010-2024, we reveal trends such as an increasing portion of reviews expressing emotional challenge, meaning, and nostalgia. We then analyze the contributions of different genres and games to the trends, in addition to reasons explicitly indicated by the reviews, and establish correlations between review scores and different player experience constructs. Taken together, our results provide novel insights into how player experiences have evolved. Methodologically, we propose and demonstrate a novel and scalable method for analyzing game reviews.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

Player Experience, PX, Game review analysis, Large Language Model, LLM

ACM Reference Format:
Supriya Dutta, Joel Oksanen, Jaakko Väkevä, Shamit Ahmed, Markus Kirjonen, and Perttu Hämäläinen. 2026. Mining Player Experience Trends From Game Reviews Using Large Language Models. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3772318.3790760

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790760

1 Introduction

Understanding player experience is fundamental to game design and research. In recent years, measuring various facets of player experiences has become possible through new validated questionnaires such as the Player Experience Inventory (PXI) [1] and Challenge Originating from Recent Gameplay Interaction Scale (CORGIS) [28]. In addition to questionnaires specifically designed for games, game research has also borrowed instruments from other fields, for example, to measure aesthetic emotions [9, 64] and intrinsic motivation [69]. However, recruiting large enough samples for questionnaire studies poses a challenge, in particular, if one would like to measure experiences *longitudinally over several years* and/or *over a wide variety of genres and games*. Simulating human participants with Large Language Models (LLMs) has been explored as a scalable data source [3, 34, 67], but it comes with problems such as reduced diversity of responses [34] and caricaturistic representation [13].

At the same time, outside academic research, thousands, or perhaps even millions of players have provided descriptions and reflections of their experiences online, in the form of game reviews. This appears as a vast data source that could enable player experience research to answer new questions, but methods have been lacking for converting such freeform texts into experience insights at scale. Various Natural Language Processing (NLP) techniques have been applied to game reviews, pioneered by Zagal et al. [83], but unfortunately, their work preceded modern LLMs and was therefore limited to more rudimentary techniques such as analyzing the frequencies of adjectives used. Follow-up work is sparse and largely focused on topic modeling and sentiment analysis (e.g., [27, 44, 53, 72]), providing relatively limited insights into player experiences. On the other hand, review-based research that goes deeper into areas such as emotional and functional challenge [16], therapeutic games [55], and intriguing first-hour experiences [14] has analyzed the data manually using approaches such as grounded theory and inductive content analysis, which are inherently limited to smaller samples of reviews.

In this paper, we seek to solve the above mentioned problems and *bridge questionnaire-based and review-based player*

experience research by mapping game reviews to player experience questionnaire responses computationally, inspired by recent advances in utilizing LLMs to predict Likert-scale questionnaire answers based on open-ended question answers [42, 43]. Our focus is on the following research questions that require a holistic and longitudinal view of multiple aspects of player experiences over several years and have not been answered by previous research:

- RQ1: How have player experiences evolved over the years, in light of the reviews? Are there any clear observable trends?
- RQ2: What are the explanations for the trends? For instance, are there particular games or genres contributing to the trends?

The core principle of our approach is simple, based on two observations: 1) Player experience questionnaire items are statements such as "This game is more than just a game to me" and the player is asked to rate their agreement with the statements, and 2) Game reviewers often narrate their experiences with similar statements. This suggests that *the semantic similarity of a review and a questionnaire item can be used as an approximation for reviewer-item agreement*, assuming that reviews are honest, on average, and review-item agreement thus indicates the reviewer's actual agreement. Naturally, a single review will not discuss every possible aspect of player experience, but following Zhu and Fang [84], we argue that reviews do often foreground at least the most important or strongly perceived aspects.

We investigate review-item semantic similarities systematically for all the items of three questionnaires commonly used in game research—PXI [1], CORGIS [28], and AESTHEMOS [64]—and a dataset of 152143 user reviews from years 2010-2024 scraped from the popular review aggregation site Metacritic[1], using a text embedding LLM to compute the similarities. This allows us to produce trend curves such as those in Figure 5, highlighting developments such as the increasing prevalence of reviews expressing emotional challenge. Furthermore, we conduct various additional analyses to shed light on the underlying reasons for the trends, and link our findings to relevant academic discourse such as a shift or expansion of focus from hedonic to eudaimonic game experiences [10, 19, 24, 25].

Contribution. We make a methodological and empirical contribution [77] in the form of the first large-scale analysis that maps game review data to established player experience constructs using LLM embeddings. Methodologically, we demonstrate that the semantic similarity of reviews and questionnaire items, calculated using embedding models, can yield interesting and meaningful insights. Empirically, we analyze player experience trends through a series of experiments:

- We uncover key trends such as the increasing prevalence of emotional challenge, meaning, nostalgia, and boredom. (Section 3, Figure 5)
- We inspect the contributions of different games and genres to the trends, surfacing games that have the most reviews relating to various experience dimensions. Although no particular game or genre appears to dominate the analysis, we demonstrate that our approach provides a new way for discovering reference games for studying and designing particular experiences. (Section 4, Figure 6, Table 2).
- We conduct an AI-assisted Qualitative Content Analysis (QCA) of the reasons the reviews provide for the trending experiences (Section 5, Table 4).
- We perform a correlation analysis to reveal which aspects of player experience dimensions are most strongly associated with game review scores (Section 6, Table 5).

Taken together, our results provide novel insights into how player experiences have evolved over time and what player experience aspects designers might want to prioritize if the goal is to ensure favourable reviews. Methodologically, we also provide a novel demonstration of the power of LLMs in mining game reviews. In future work, our methodology should be applicable to both other questionnaires and other data sources such as social media posts. Python source code and data is available at https://version.aalto.fi/gitlab/gameresearch/PXTrends.

2 Background and Related work

Below, we review the related background and previous research, divided into three main parts. First, we briefly survey various player experience frameworks to ground our choice of questionnaires. Second, we discuss the use of LLMs in text analysis and psychological assessment. Finally, we review research in our specific domain of analyzing game reviews, highlighting the heretofore untapped opportunity of applying LLM-based computational psychological assessment to such data.

2.1 Understanding and Measuring Player Experience

Player experience is a broad concept that can be discussed on (socio-)psychological, behavioural, and physiological levels [76]. Correspondingly, there exists a multitude of approaches for measuring and assessing player experience. On a behavioural level, one might measure variables such as the frequency and length of play sessions and how they evolve over time, whereas common physiological measurements include skin conductance and tracking facial expressions via electromyography or computer vision [40, 61]. Of course, both physiological and behavioural measures reflect the user's internal state such as emotions and motivations and are therefore closely connected to the socio-physiological level.

In this paper, our focus is on player experience on a psychological level, measured through verbal self-reports. Many questionnaires have been proposed and utilized for this, both game-specific and from other fields, based on different assumptions and psychological models such as Self-Determination Theory and Flow [76], dividing player experience into key

[1] https://www.metacritic.com/

components such as Competence, Autonomy, Relatedness, and Curiosity, to name a few [72]. Below, we explain the questionnaires used in this paper in more detail.

It should be noted that our choice of questionnaires is by no means complete, or a statement about the value of the included questionnaires in relation to questionnaires we excluded, such as the recent BANGS scale [6], which expands the study of psychological need satisfaction in games to also include need frustration. Given the paper length limit, we simply could not fit in more questionnaires. Our selected three questionnaires are used actively in games and HCI research, each cited by over 100 articles that include the word "game" in the past 5 years, according to Google Scholar. The questionnaires are also fairly recent (published in 2017 or later) and originate from different theoretical backgrounds, allowing us to cover a broad spectrum of experience constructs. The familiarity of the authors with the instruments was also considered as a selection criterion, as we needed to be able to evaluate and reflect on whether the embedding similarity numbers and other results make sense. As our approach is straightforward to implement, we hope others will expand our analyses to other questionnaires and datasets in future work.

2.1.1 Player Experience Inventory (PXI) [1, 38]. PXI is a tool for assessing player experience across various game genres and gamified applications. Rooted in Means-End Theory and the Mechanics, Dynamics, and Aesthetics (MDA) framework, PXI evaluates player experience on two levels. First, it examines Functional Consequences, which refer to the immediate and tangible effects of game design choices. These comprise five constructs: Ease of Control, Goals and Rules, Challenge, Progress Feedback and Audiovisual Appeal. Each of these constructs consists of three questionnaire items. For instance, for ease of control, the three items are "It was easy to know how to perform actions in the game", "The actions to control the game were clear to me", "I thought the game was easy to control". Second, PXI assesses Psychosocial Consequences, capturing the emotional responses that emerge as an indirect result of those design choices. These comprise five constructs as well: Meaning, Curiosity, Mastery, Immersion, and Autonomy. PXI has been validated and evaluated across seven studies with 529 participants [1]. The findings confirm the scale's intended structure and display evidence for both discriminant and convergent validity.

2.1.2 The Challenge Originating from Recent Gameplay Interaction Scale (CORGIS) [28]. CORGIS is designed to assess four different dimensions of perceived challenge. The conceptual foundation of CORGIS is based on literature on game user research, design, and AI, along with qualitative insights from interviews with both researchers and players. CORGIS identifies four primary dimensions of challenge, derived through Exploratory Factor Analysis (EFA): Performative, Emotional, Cognitive, and Decision-making challenge. The scale has been validated in with nearly 1,000 players, demonstrating strong construct validity [28]. CORGIS comprises 30 items, distributed across the four challenge components. Examples of the items include "Succeeding in the game required much planning" (Cognitive challenge), "The game made me think about real-life issues" (Emotional challenge), "I had to react quickly when playing the game" (Performative challenge), and "The game made me think hard about my decisions" (Decision-making challenge).

2.1.3 Aesthetic Emotions Scale (AESTHEMOS) [64]. AESTHEMOS is not a questionnaire specific to games, but it has been used, e.g., to study games as art experiences [9]. In this paper, we adopt AESTHEMOS to provide more resolution on the emotional aspects of player experience. AESTHEMOS is based on theoretical perspectives on aesthetic emotions and a review of existing measures across various domains, including music, literature, film, painting, advertising, design, and architecture. The questionnaire comprises 21 subscales, each consisting of two items. AESTHEMOS captures a broad spectrum of emotions, namely, Prototypical aesthetic emotions (nostalgia, feeling of beauty, awe, being moved, enchantment, fascination), Epistemic emotions (surprise, insight, intellectual challenge, interest), Pleasing emotions (vitality, humour, joy, energy, relaxation) and Negative emotions(feeling of ugliness, sadness, boredom, anger, confusion, uneasiness). Examples of items include "I found it beautiful", "Made me curious", "Calmed me", "Delighted me", "Felt a sudden insight".

2.2 LLMs in Text Analysis

2.2.1 Two Types of LLMs. There exist two main types of LLMs: embedding models and autoregressive generative models. Embedding models [46, 52, 60] map a piece of text, such as a game review, into a high-dimensional vector space, such that the distance between vectors for semantically similar texts is small, typically measured through cosine similarity. We utilize this capability to calculate similarities between reviews and questionnaire items.

On the other hand, autoregressive generative models, such as GPT-3, GPT-4, and Llama [2, 12, 29], generate text by predicting one token (word or a word piece) at a time to continue an initial seed text, i.e., a prompt. Each predicted token is then added to the prompt to inform the next predicted ones. In a chat-based setting, the prompt is formatted as a dialogue between the user and an AI assistant, allowing generative LLMs to be instructed to perform various text analysis tasks, such as producing structured summaries of long texts (e.g., [45]).

2.2.2 Qualitative Text Analysis. Recent research has increasingly explored the role of LLMs in qualitative data coding, a key method for analyzing textual qualitative data where researchers label segments of text with short descriptive *codes* [50]. For example, Hämäläinen et al. [34] demonstrated that LLM coding is possible using in-context learning [12], while Dai et al. [23] introduced a human-LLM collaboration framework that uses few-shot examples and iterative refinement to align model outputs with researcher intentions. A number of studies have explored the utilization of LLMs in qualitative

research within various domains, from media analysis [30] to phenomenological research [35].

Although LLMs can accelerate coding workflows and support the emergence of new insights, several studies have cautioned against relying solely on automated outputs. LLM-generated codes may introduce systematic biases or lead to superficial interpretations if not carefully moderated by human researchers [4, 66]. As a result, there is a growing consensus that LLMs should augment, rather than replace, human analysis, for example, through grounding the automated analysis in an initial set of manually coded examples. In our work, we utilise LLMs for deductive coding which is more straightforward than inductive coding, based on human-defined codebooks based on samples of manually coded data.

2.2.3 Computational Psychological Assessment. Our work capitalizes on the recent stream of psychological research on predicting closed-ended responses (e.g., Likert-scale numerical ratings) based on open-ended responses (e.g., social media posts, descriptions of one's mental state) [41–43, 54, 62]. As Kjell et al. [41] pointed out, it is more natural for people to give an open-ended answer to questions such as "How are you?", even though psychological assessment has traditionally utilized closed-ended answers. Fortunately, with modern computational text analysis methods, it is becoming possible to predict closed-ended answers based on open-ended ones.

Such prediction approaches were initially based on embeddings and used additional machine learning components. Kjell et al. [41] used linear regression with individual word embedding vectors as the input features, and their follow-up word showed that embedding full texts using the BERT LLM improves the results, approaching the theoretical upper limits in accuracy [43]. Hitsuwari et al. [36] also utilized BERT but by finetuning it. The drawback of these approaches is that they require ground truth data of paired open-ended and closed-ended responses for fitting the regression parameters and finetuning the LLM. Later work has explored prompting-based approaches where the open-ended response or reference text is simply included in an LLM prompt that asks the LLM to output the closed-ended response [54]. Such prompting-based does not require ground truth data for parameter fitting, but on the other hand, can be much more computationally intensive, as we discuss in Section 3.2.3. There also exists research combining LLM prompting with custom machine learning, by first prompting the LLM to provide intermediate numerical ratings that are then used as input features to an additional predictive model [37, 62].

Our approach is directly inspired by Atari et al. [5]. Similar to their work, we estimate the agreement of questionnaire items and open-ended texts simply through embedding similarity, which requires no paired ground truth data. We extend their work by 1) utilizing the approach to investigate trends in large-scale longitudinal data, and 2) applying additional thresholding to reduce noise, which turned out to be essential in uncovering trends from our data.

Notably, none of the existing work on computational psychological assessment has been done in the domain of game research, instead focusing on areas such as well-being and mental health [41, 43], personality [54], and life satisfaction [37].

2.3 Game Review Research

Researchers have used different methods to analyze and understand game reviews written by players. In the influential work by Zagal et al. [83] using traditional (pre-LLM) Natural Language Processing (NLP) techniques, common themes were identified, including pacing (how often events occur in the game), complexity (how different parts of the game interact), cognitive accessibility (how easy the game is to understand), scope (the range of possible actions), demands (the challenges the game presents to the player), and impact (the emotions the game evokes).

The degree to which a topic is discussed in a game review depends on the type of game being reviewed. For example, Cole et al. [17] conducted a qualitative analysis of online reviews for 14 games, including both what they referred to as *core* and *avant-garde* titles. The authors defined core games as commercially successful blockbuster games and avant-garde games as those praised for pushing the boundaries of the medium. Their study found that reviews of core games frequently discussed gameplay challenges, a variety of mechanical and technical aspects—like character design, lighting, and graphics—and emotions often associated with action films. In contrast, reviews of avant-garde games focused more on emotional challenge, the quality of the gameplay experience, a broader range of emotional experiences—including those involving reflection and contemplation—and the emotional impact of sound and visuals.

Studies using NLP topic modeling and sentiment analysis in game reviews have identified key themes such as narrative, achievement, social interaction, visual appeal, social influence, accessories, and overall player experience [72]. Reviews about the first hours of gameplay have been found to often discuss players' expectations—shaped by past experiences or gaming communities—followed by their in-game experiences, which influence their decision to continue playing [14]. Cultural background also plays a role in how players write reviews. A study by Pan *et al.* [53] on Souls-like games, which analyzed reviews in English, Chinese, and Russian, found that Chinese and Russian reviews focused more on combat mechanics compared to English reviews. Additionally, neutral emotions were more commonly expressed in English and Russian reviews than in Chinese reviews. Topic modeling has also been used to analyze online reviews to identify the challenges players face during gameplay and to improve the overall experience in multiplayer games [81, 82].

In studying review data, a key choice is between a focus on user reviews vs. expert reviews which differ in content. For instance, user reviews tend to describe experiences while experts focus on factual descriptions [63], which is a good fit

for our focus on player experience. On the other hand, instead of providing an honest appraisal of a game, user reviews may express other things such a disagreement with the game's developer [63]. We chose to focus on user reviews primarily due to the larger amounts of both reviews and their writers, which should result in more reliable longitudinal trend data and more diverse perspectives.

In the domain of game review research, the work closest to ours is by Lankes and Stöckl [45], who provide the only previous instance of combining review data and a player experience questionnaire (PXI), albeit for a different purpose. Instead of mapping reviews to PXI responses, they prompted GPT-3 with reviews and PXI-based questions such as "What was meaningful to players while playing the game?", in order to produce bullet-point summaries about the player experience. Our embedding-based approach answers different research questions and is also more scalable than a prompt-based approach, as detailed in Section 3.2.3.

3 Experiment 1: Visualizing Trends

To begin our exploration, this experiment visualizes player experience trends based on the PXI, CORGIS, and AESTHEMOS questionnaires and Metacritic user reviews from 2010-2024.

3.1 Data

We used user reviews from Metacritic, a platform that consolidates game details and reviews across multiple platforms. While Metacritic includes both critic and user reviews, we focused on user reviews as they provide a larger sample and might offer access to more raw, unfiltered player emotions and experiences. The dataset was scraped using Selenium and Python, capturing all reviews available online. In total, the dataset comprises 152,143 user reviews spanning 9,107 unique games, covering multiple platforms and genres, with timestamps ranging from November 17, 2010, to August 1st, 2024. The review length distribution is visualized in Figure 1. Overall, the user reviews are fairly short, with a median length slightly over 200 characters. The review length has also been gradually declining, especially since 2017.

3.2 Method

As explained in the introduction, we use the semantic similarity between a review and a questionnaire item as a proxy for the reviewer's agreement with the item. This is in no way perfect and requires some extra processing steps to reduce noise, as illustrated in Figure 2. Below, we first explain the approach and then elaborate on the rationale for choosing the embedding-based approach over the alternative of prompting a text-generating LLM for synthetic questionnaire responses.

3.2.1 Calculating Semantic Similarity. Semantic similarity was measured by first transforming each review and questionnaire item into a 3072-dimensional embedding vector using the multilingual OpenAI text-embedding-3-large model [52], which at the time of writing had a Massive Text Embedding Benchmark (MTEB) average score of 64.52 [31]. Then, semantic similarity between each questionnaire item q and review r was calculated as the cosine similarity:

$$similarity(q, r) = \frac{\mathbf{x}_q^T \mathbf{x}_r}{|\mathbf{x}_q||\mathbf{x}_r|}, \tag{1}$$

where $\mathbf{x}_q$ and $\mathbf{x}_r$ denote the embedding vectors for the items and reviews, respectively. Table 1 shows illustrative examples of the similarities of different reviews and questionnaire items.

3.2.2 From Similarities to Percentages of Reviews. Initially, we tried visualizing trends using yearly means and medians of the similarity values, as shown in Figure 3. However, it appears that the data is so noisy that no trends are visible this way, even when averaging over the items of each subscale. There are many contributors to such noise. In particular, the quality and length of the user-written reviews has a high variance, some reviews including rather incoherent text such as "this Xpac is a waste of money plain and simple! only real reason i got HOT waz for the claimed PVP matchmaker that waz released late and is 100X worse then Mechwarrior:online's **** thats saying alot (-_-) then WvW new cliff fall simulator that is EZ'er then hell...". Some reviews do not express anything about player experience, and many reviews only comment only on some experience dimensions, resulting in non-informative similarity scores for the other dimensions.

To solve the issue and remove noise, we *count and plot the yearly percentages of all reviews where the similarity is above a threshold value.* In effect, the thresholding performs a binary classification of the reviews into those that highlight a particular experience dimension and those that do not highlight it. This makes sense if one makes the following two assumptions: 1) If a player experiences, say, emotional challenge particularly strongly, they feel the need to comment on it in the review, resulting in a high similarity with items measuring emotional challenge. 2) Otherwise, the review might not comment on emotional challenge at all, focusing on other things, resulting in non-informative similarity that is typically low, but has high variance depending on things like the usage of words related to emotional challenge.

Specifically, for each subscale s and each review r, we first calculate the review-subscale similarity as the average of similarities over the subscale items:

$$subscale\_sim(r, s) = \frac{1}{|Q_s|} \sum_{q \in Q_s} similarity(q, r), \tag{2}$$

where Q_s is the set of all items for the subscale s. Assuming that $similarity(q, r)$ is proportional to how the writer of review r would rate their agreement with item q, the averaging corresponds to calculating a subscale score by additively combining the individual Likert-scale responses, which is how PXI, CORGIS, and AESTHEMOS are designed. We calculate an average instead of a sum to make the result invariant of the number of subscale items. Note that before averaging, we

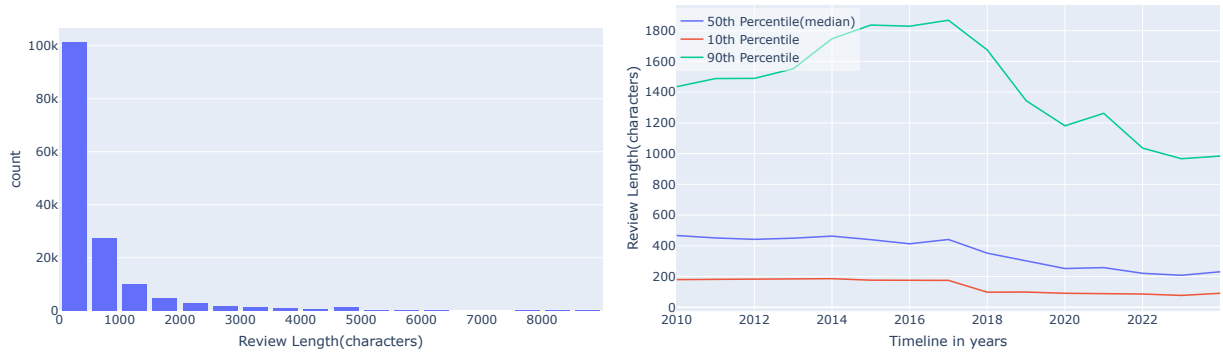


Figure 1: Left: Histogram of review lengths. Right: Evolution of review length over time. Overall, review lengths have been slightly declining, especially since 2017.

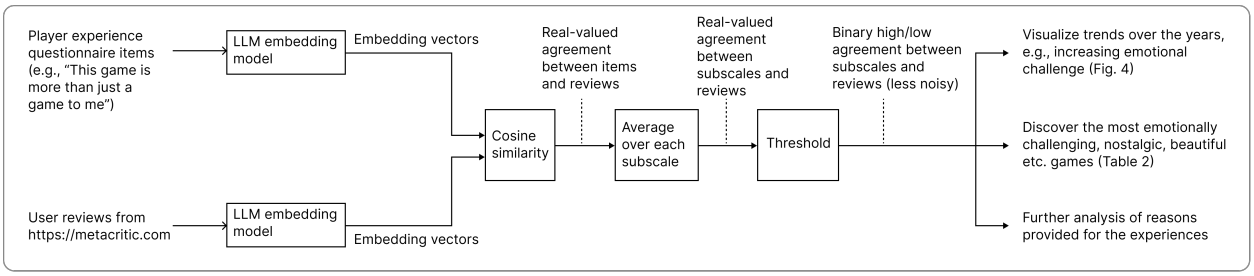


Figure 2: Overview of our processing pipeline. We estimate the agreement of reviews with questionnaire items via semantic similarity (cosine similarity of LLM embedding vectors). The raw similarities are noisy, which we mitigate by averaging the similarities over each subscale and thresholding to obtain binary review-subscale agreements.

first normalize the similarities to the range 0...1, separately for each questionnaire.

Then, for each subscale s and year y, we calculate the percentage of reviews with the similarity over a threshold T:

$$percentage(y, s) = \frac{100}{|R_y|} \sum_{r \in R_y} \mathbf{1}_{subscale\_sim(r,s)>T}, \quad (3)$$

where $\mathbf{1}$ denotes the indicator function and R_y is the set of all reviews for year y.

The percentage plots in Figure 4a and 4b show a trend of reviews increasingly expressing emotional challenge, in contrast to the means and medians of Figure 3. The threshold value adjusts the tradeoff between false positives and false negatives: A too low value results in including non-relevant reviews (false positives) and produces constant or nearly constant yearly percentages with no visible trends. A higher threshold reduces the number of false positives, emphasizing the trend, but as shown in Figure 4c, a too high value can result in only a few reviews per year, making the percentage calculations noisy. Through trial and error, we selected a threshold $T = 0.6$ for PXI and CORGIS, and $T = 0.45$ for AESTHEMOS, aiming for clearly visible trends with minimal noise. Note that the per-questionnaire thresholds mean that the prominence of the visualized trends are only comparable within the same questionnaire and not across questionnaires.

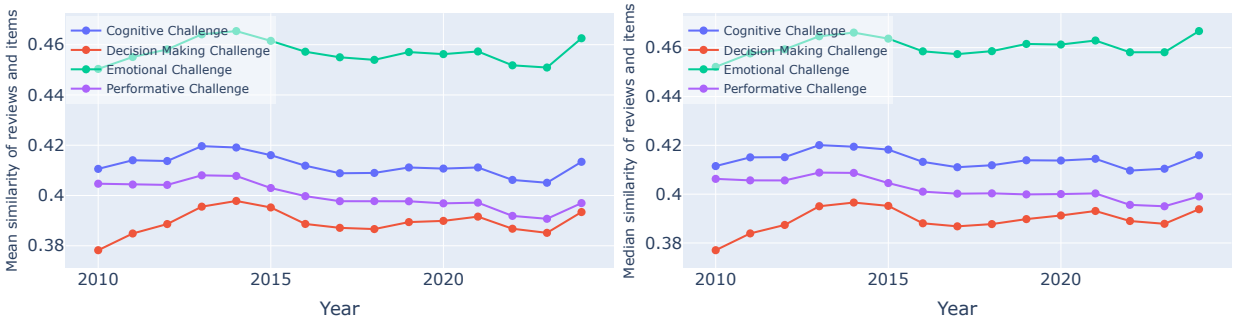
3.2.3 Embedding vs. Prompting. An alternative to using an embedding LLM would be to use a text generation LLM such as ChatGPT with a prompt such as *"You are a game reviewer who has written the following review <REVIEW>. Please rate your agreement with the statement <QUESTIONNAIRE ITEM> on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree)".*

We chose the embedding-based approach primarily due to its better scalability to large datasets. The prompt-based approach requires a separate prompt for each review-item pair, i.e., a total of NM queries to a LLM, where N is the total amount of reviews and M is the total amount of items. In contrast, the *embedding vectors can be queried separately for the reviews and items*, resulting in only $N + M$ queries, and for each review-item pair, one only needs to calculate the cosine similarity which does not require access to any paid LLM API and is fast even on a typical laptop computer.

In our case, $N = 152143$ and $M = 102$, i.e., the prompt-based approach would require over 100 times more LLM queries. Furthermore, the per-token API cost of embedding models is typically lower. For instance, at the time of writing this,

Review	Similarity	Questionnaire item
"Its so pretty"	0.50	I found it beautiful
"A breathtaking work of art, that t'ill this day, never ceases to amaze me."	0.39	I found it beautiful
"It was an awesome game when I first played this game"	0.28	I found it beautiful
"This is an excellent, nuanced game. If you enjoy a complex strategy game that actually works, I would recommend this game. The game requires strategy in order to succeed: you must build up various supply lines of manufactured goods via conquering and collecting resources from various strategic points. There is also a well-thought-out technology tree that requires a well-thought-out plan."	0.46	Succeeding in the game required much planning
"My first strategy. And, i think, the best. There is no any trash in this game. All, what you got, you use. The best in this game is economy part. You nothing if your army don't have secure rear."	0.43	Succeeding in the game required much planning
"Wow, this game was incredible back in 1995 when I was a sophomore in college. My friends and I were in awe even with its DOS talking installer. The game was rad. Sure, there were bugs, multiplayer were iffy, etc."	0.28	Succeeding in the game required much planning

Table 1: Examples of the semantic similarity between reviews and questionnaire items, calculated as the cosine similarity of the review and item embedding vectors.



(a) Mean similarity **(b) Median similarity**

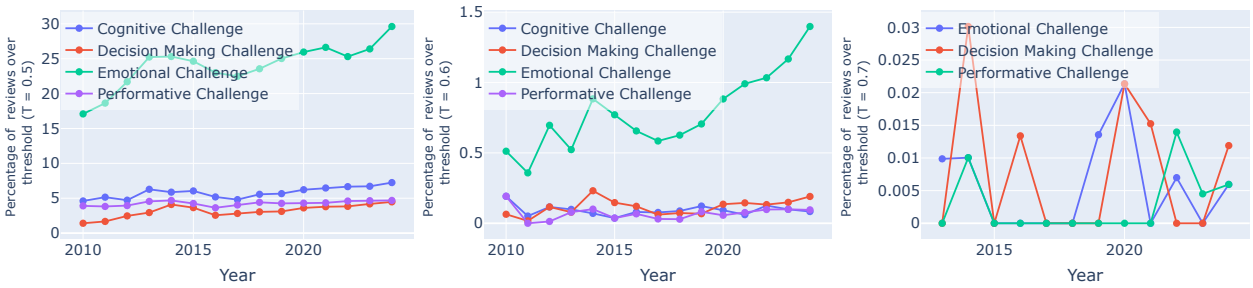
Figure 3: Mean and median similarity of reviews and questionnaire items, visualized for the four subscales of CORGIS. No clear trends can be observed.

the OpenAI text-embedding-3-large model costs $0.13 per 1M tokens, compared to $2.50 for the GPT-4o text generation model and $0.15 for the cheaper but less capable GPT-4o-mini.

Another way to improve quality could be to map review embedding vectors to questionnaire responses via linear regression similar to Kjell et al. [41, 43], but this is not possible as the Metacritic data does not contain the ground truth questionnaire responses required for fitting the regression model.

3.2.4 Full Reviews vs. Individual Sentences. Our unit of analysis is full review texts instead of individual sentences. This is motivated by the short length of the Metacritic user reviews which makes them well-suited for embedding. The OpenAI Text Embedding 3 Large model's maximum number of input tokens is 8192, which none of the reviews violate. Furthermore, embedding and thresholding single sentences would cause more false positives and would be more sensitive to language evolution such as a particular catch phrase becoming popular. Considering that many user reviews feature grammar and punctuation mistakes, the reliability of sentence extraction is also questionable.

3.2.5 Construct Validity. In summary, we use the cosine similarity of the review and item embeddings as an approximation for review-item agreement, assuming that this, in turn, reflects actual reviewer agreement. We then detect highly agreeing



(a) Similarity threshold: $T = 0.5$ **(b) Similarity threshold:** $T = 0.6$ **(c) Similarity threshold:** $T = 0.7$

Figure 4: Percentages of reviews with high agreement with questionnaire items, visualized for the four CORGIS subscales. Crucially, compared to the means and medians of Figure 3, the percentage visualization surfaces a trend in Emotional Challenge. However, as shown on the right, a too high threshold value results in so few reviews that calculating percentages becomes brittle.

reviews using the thresholding. For a validation of the approach based on a sample of reviews coded manually by three human coders, see Appendix A. Based on the validation, the cosine similarity indeed appears a useful approximation, with a moderate-to-strong correlation with human-rated agreement. Furthermore, the validation confirms our motivation for the thresholding: Only a minority of reviews agree with a specific item, with most review-item pairs rated as neutral (no agreement or disagreement). The neutral reviews also exhibit a large variance in cosine similarity, explaining why trends are not visible in the mean and median graphs of Figure 3a and Figure 3b.

3.3 Results

The trends for all three questionnaires are visualized in Figure 5. For CORGIS, there is a gradual increase of Emotional Challenge. In the PXI dataset, Audiovisual Appeal, Meaning, and Mastery exhibit a consistent upward trend. The AESTHEMOS data reveals gradual increases in Boredom, Beauty/liking, Nostalgia, Joy, Humour. Furthermore, AESTHEMOS Interest and Surprise have also been increasing, but this appears less reliable due to the high noise.

Of course, simply observing the trends is of little value without understanding the underlying reasons. As game tools and technology have progressed, it has become easier to produce visually appealing games, which is probably a major contributor to the audiovisual appeal and beauty trends. However, for other trends such as emotional challenge and nostalgia, the reasons are less obvious. Thus, the rest of this paper continues with three further experiments. Section 4 first investigates the contribution of individual genres and games to the trends. Section 5 then proceeds with a more detailed analysis of the reasons the reviews state for experiencing things like nostalgia. Finally, and Section 6 analyses the association of different experience constructs with high or low game review scores.

4 Experiment 2: Contributions of Individual Games and Genres

To understand the underlying reasons for the trends highlighted by Experiment 1, this experiment explores the impact of individual games and genres. The Metacritic review data includes zero or more genre tags for each review. For readability of the results, we focus on the following 10 trends that were both prevalent in Experiment 1 and that we find the most interesting as from the research and design perspective: Emotional challenge, Audiovisual appeal, Meaning, Mastery, Nostalgia, Beauty/liking, Joy, Humour, Surprise, and Boredom.

As the primary method of investigation, we utilize stacked area charts: For each year, we calculate the percentages of reviews for each genre and game included in the trends, and generate corresponding stacked area charts of each trend so that the top of the chart reproduces the original trend curve in Figure 5 and the areas display the relative contributions of the genres and games. To maintain the readability of the charts, we only show the top 10 contributors separately and group the others in an "Others" category. The top 10 contributors were obtained by sorting the genres and games by the total number of reviews with the similarity over the threshold.

4.1 Experiment 2 Results

4.1.1 Trend Contributions. Generally, the stacked are charts reveal no single or a few clearly dominating games or genres. Hence, we only include the visualization for Emotional Challenge as an example, in Figure 6. Similar visualizations for the rest of the trends can be found in the Appendix D.

Based on Figure 6, the point-and-click genre was a notable contributor to emotionally challenging experiences from 2012 to 2016 but declined thereafter. There are no clear changes in the other genres, except perhaps the increasing contribution of survival games since 2018. Other genres such as adventure, action-adventure, and first-person adventure have been steadily contributing throughout the years.

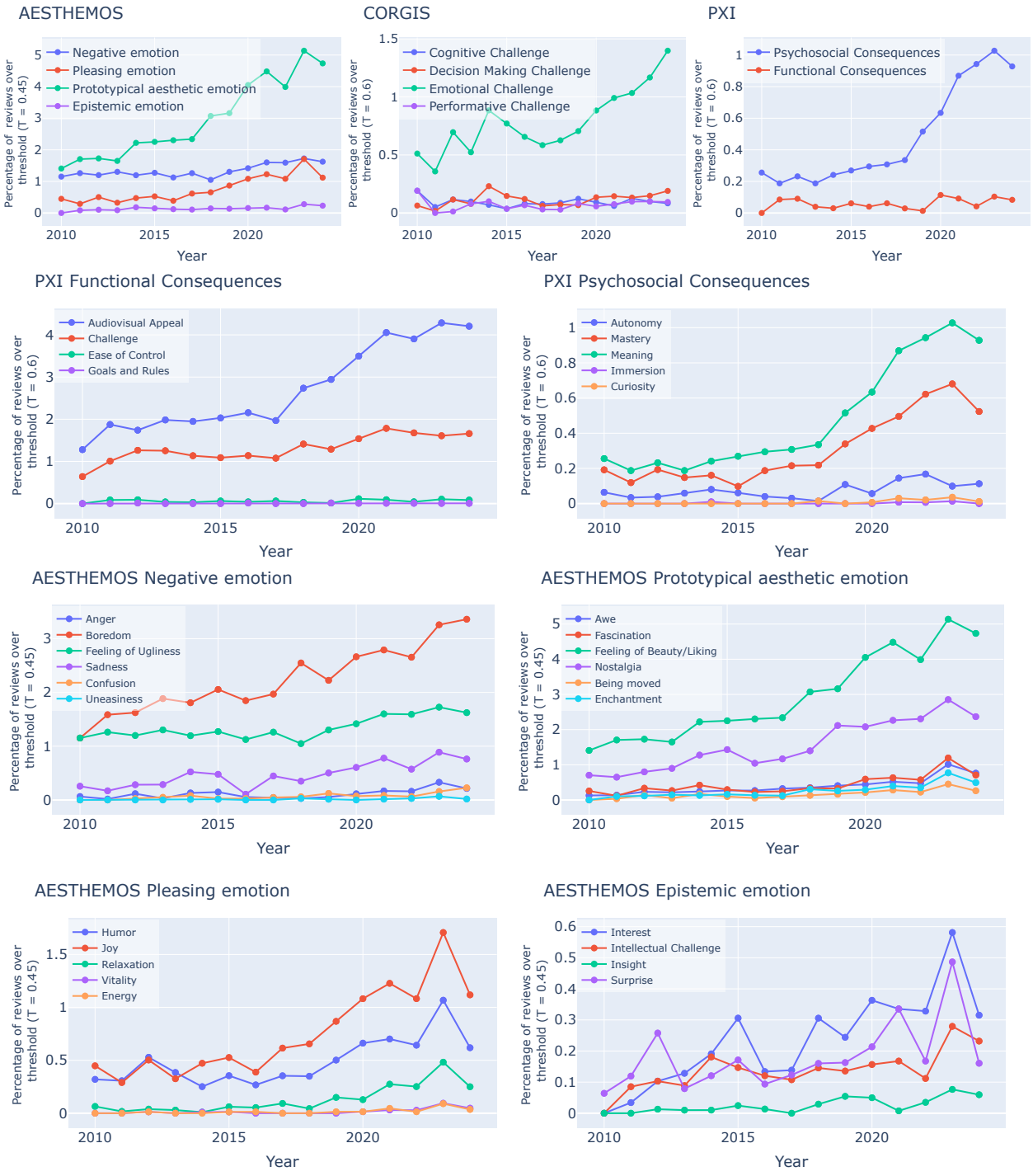


Figure 5: Experience trend curves. The x-axes denote the year a review was written. The y-axes denote the percentage of reviews agreeing with the questionnaire items, calculated according to Equation (3). The top row shows trends using the high-level experience categories while the other rows show breakdowns on the level of individual constructs. CORGIS measures only the 4 constructs shown on the top row.

The visualization of the individual games contributing most to Emotional Challenge, on the right in Figure 6, reveals no single clearly dominating game, except perhaps the 2012 peak for The Walking Dead: Episode 5 - No Time Left [G52]. The Walking Dead is categorized in the point-and-click genre, and it might be that the game inspired many others, providing at least a partial explanation for the point-and-click genre's dominance in 2012-2016.

4.1.2 Auxiliary Finding: A Method for Game Discovery. Even though the trends appear to be a compound effect of multiple games, with no clearly dominating ones, the top 10 games do feature many games already discussed in emotional challenge research. Thus, it appears that this kind of sorting of games can provide *a new way to discover representative games to play, analyze and discuss*. To facilitate this, we have collected all the top 10 games for all the trends into Table 2. For each game, the table includes a clickable link to the ludography (e.g., [G37]), which then provides links to more information. To filter out chance findings, the Table 2 only includes games with at least 5 reviews with the review-subscale similarity over the threshold.

5 Experiment 3: Qualitative Content Analysis of Reasons Provided

As the preceding experiments surfaced limited insights on the reasons underlying the trends, we conducted a Qualitative Content Analysis (QCA) into to further investigate a selection of the trends: Emotional Challenge, Boredom, Meaning, and Nostalgia. Emotional Challenge, Meaning, and Nostalgia were included based on our own research interests and because they are related to the broader concept of eudaimonic gaming experiences, which is a topic that has received increased attention in the game research literature in recent years [24, 25, 57]. Boredom was further included to extend the analysis to both positive and negative player experiences. Investigating all the 10 trends visualized above was not considered feasible in terms of paper length and the resources of the research team.

5.1 Experiment 3 Methods

The analysis started with manual inductive coding and codebook creation based on a random sample of the data. This was followed by large-scale LLM-assisted deductive coding of the full dataset. We repeated the following steps for each of the four trends we investigated:

(1) A random sample of 200 reviews with cosine similarities over the threshold was composed.
(2) Two independent coders (two of the authors) coded the sets of 200 reviews inductively, creating their own codebooks for the reasons the reviews stated for the investigated experience.
(3) All the authors discussed the coding results and collaboratively compiled a final codebook.
(4) The same two coders re-coded the 200 reviews deductively, using the final codebook.
(5) The OpenAI GPT-4-Turbo and GPT-4o LLMs were employed to deductively code the full datasets for each 4 trends, i.e., all the reviews with cosine similarity over the threshold. This resulted in a total of 8665 coded reviews (1,364 for Emotional Challenge, 3,785 for Boredom, 923 for Meaning, and 2,593 for Nostalgia). The LLM was given the same deductive coding instructions as the human coders.
(6) The LLM coding results were validated by calculating both human-LLM and human-human inter-coder agreements using the 200 manually coded reviews.
(7) The coding results were compiled to a table sorted by code frequencies, therefore surfacing the most prominent reasons.
(8) The evolution of code prevalence over time was visualized using stacked area charts similar to Figure 6 above.

Below, we elaborate on the methodological details and rationale.

5.1.1 Rationale for LLM use. Although the reliability of LLMs for qualitative data analysis can be questioned [4, 66], we consider our LLM usage valid for a number of reasons. First, we only utilize LLMs for deductive coding. Compared to inductive coding which can require considerable interpretation and reflection from the researcher, deductive coding is more straightforward—essentially, it is a text classification task, in which LLMs are known to excel [73, 85] and this has also been validated specifically in the context of qualitative text analysis [15]. Second, our manual analysis of the sets of 200 reviews by two independent coders allows us to validate the results via the LLM-human inter-coder agreement. Finally, the large number of total reviews was not feasible to analyze manually. Even though the manual analysis of the samples of 200 reviews (800 in total) does give us approximate code frequencies, visualizing the evolution of code frequencies over time would require analyzing separate samples of reviews for every year. All things considered, we believe the potential LLM-induced analysis inaccuracy is an acceptable trade-off.

5.1.2 Inductive Coding. For each of the four trends investigated, the coders inductively created their own codebooks such that the codes denote reasons expressed by the reviews for experiencing the construct of interest. The instructions given to the coders reminded them about the precise operationalization of the investigated constructs by listing the corresponding questionnaire items. For instance, the instructions for PXI Meaning were:

> Your task is to help in analyzing the reasons for experiencing meaning in games, based on a dataset of game reviews. Here, experiencing meaning is conceptualized as an agreement with the following statements:
>
> Playing the game was meaningful to me.
>
> The game felt relevant to me.

No.	Nostalgia (AESTHEMOS)	Feeling of Beauty (AESTHEMOS)	Joy (AESTHEMOS)	Humour (AESTHEMOS)	Surprise (AESTHEMOS)
1.	Before Your Eyes (8.5)[G62]	RiME (7.8)[G55]	Party Animals (7.0) [G46]	High on Life (7.7) [G49]	no games with more than 5 reviews for this category
2.	To the Moon (8.9) [G17]	The Artful Escape (7.5) [G4]	Before Your Eyes (8.5) [G62]	The Stanley Parable: Ultra Deluxe (8.3) [G20]	
3.	Emily is Away (7.2) [G30]	Darkarta: A Broken Heart's Quest (8.8) [G60]	The Gunk (6.8) [G63]	Save Farty (8.7) [G57]	
4.	Teenage Mutant Ninja Turtles: Shredder's Revenge (8.1) [G59]	GRIS (8.2) [G37]	Returnal (7.4) [G29]	Trover Saves the Universe (8.1) [G48]	
5.	Unravel (8.0) [G8]	Fort Solis (6.7) [G16]	Super Mario Bros. Wonder (9.0) [G36]	The Henry Stickmin Collection (8.7) [G42]	
6.	Hypnospace Outlaw (7.3) [G54]	ABZU (7.5) [G22]	Tchia (6.6) [G1]	South Park: The Stick of Truth (8.5) [G39]	
7.	The Beginner's Guide (7.5) [G15]	Unravel (8.0) [G8]	Princess Peach: Showtime! (6.8) [G25]	DLC Quest (6.5) [G24]	
8.	System Shock (9.0) [G35]	N.E.R.O.: Nothing Ever Remains Obscure (7.5) [G64]		Jazzpunk (7.8) [G34]	
9.	Spyro Reignited Trilogy (8.3) [G58]	Planet of Lana (7.9) [G61]		Donut County (7.1) [G5]	
10.	Firewatch (7.3) [G6]	Cocoon (8.0) [G21]		Deadpool (7.4) [G28]	

No.	Boredom (AESTHEMOS)	Emotional Challenge (CORGIS)	Audio Visual Appeal (PXI)	Mastery (PXI)	Meaning (PXI)
1.	2064: Read Only Memories (3.5) [G32]	Before Your Eyes (8.5) [G62]	Ravenlok (7.5) [G7]	Ghostrunner (7.9) [G40]	Before Your Eyes (8.5) [G62]
2.	The Walking Dead: The Telltale Series - A New Frontier (6.6) [G53]	Firewatch (7.3) [G6]	The Gunk (6.8) [G63]		drowning (Rapture) (9.0) [G2]
3.	Fallout: New Vegas - Honest Hearts (7.0) [G38]	Detroit: Become Human (8.7) [G43]	The Artful Escape (7.5) [G4]		The Beginner's Guide (7.5) [G15]
4.	Gone Home (5.5) [G19]	The Cat Lady(8.7) [G27]	The Pathless (8.0) [G23]		Tell Me Why (5.8) [G14]
5.	Tell Me Why (5.8) [G14]	Life is Strange (8.6) [G13]	Owlboy (7.6) [G9]		Prototype 2 (7.1) [G45]
6.	The Vanishing of Ethan Carter (7.8) [G56]	The Walking Dead: Episode 5 - No Time Left (8.7) [G52]	drowning (Rapture) (9.0) [G2]		Journey of the Broken Circle (7.1) [G31]
7.	Mirror's Edge Catalyst (5.9) [G10]	The Walking Dead: A Telltale Games Series (8.8) [G51]	Fort Solis (6.7) [G16]		Firewatch (7.3) [G6]
8.	Tales of Zestiria (7.2) [G3]	Heavy Rain (8.5) [G44]	GhostWire: Tokyo (8.3) [G50]		Disney Cory in the House (9.3) [G26]
9.	Firefall (5.9) [G47]	LISA (8.6) [G12]	RiME (7.8) [G55]		
10.	Bound (2016) (7.0) [G41]	Road 96 (7.7) [G11]	Ori and the Blind Forest: Definitive Edition (8.3) [G33]		

Table 2: The top 10 games contributing to the trends. To learn more about a game, click on the game's reference such as [G37] to view the ludography which will provide a link to the game's Steam store page or other extra information. The numbers in parentheses after each game are the Metacritic user score averages.

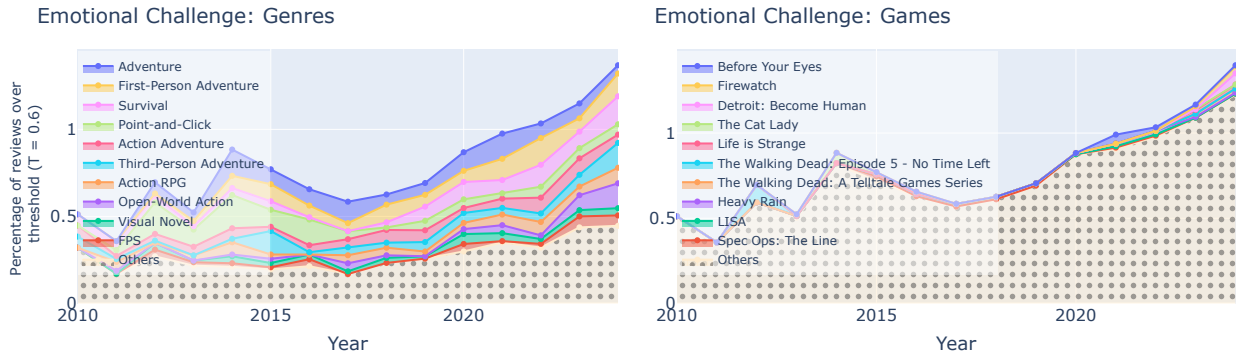


Figure 6: A breakdown of the emotional challenge trend curve in Figure 5 into the top-10 contributing genres and games. The Point-and-click genre contributed notably between 2011 and 2017 but declined since. On the other hand, the Survival genre has contributed increasingly since 2018. Regarding the top-10 games, no single game appears to make a major contribution to the trend, except perhaps The Walking Dead: Episode 5 - No Time Left [G52] in 2012.

> Playing this game was valuable to me.
>
> You should code the reviews inductively using one or more codes for each review, to indicate the reasons for the reviewer's experience of meaning.

In addition to the codes defined inductively by each coder, the codes "Not relevant" and "No Reason" were utilized for cases where the coder did not consider the review as relevant (i.e., a false-positive result from the embedding-based thresholding) or no reason was provided or could not be inferred. Regarding the "Not relevant" code, the coders required clarification about whether or not they should count a review as relevant if it agreed with even a single item. Considering that the questionnaires are designed for additively combining the Likert-scale responses, we clarified the instructions as: "Precisely, we define <construct> as the average of the agreements with each individual statement. Thus, <construct> can manifest as a strong agreement with some statements, or a moderate agreement with many statements."

5.1.3 Deductive coding. The deductive coding used the same instructions as above, augmented with the final codebooks. The same instructions were given to both the human coders and the LLM, thus allowing a direct comparison of LLM and human code quality, as detailed below. The full set of instructions are included in Appendix C.

In order to validate the quality of the LLM coding, code quality metrics were used to compare the pairwise agreement between each of the two human coders (denoted henceforth as C1 and C2) as well as the LLM for the sets of 200 reviews annotated by the human coders. Given that the codes were not mutually exclusive, i.e., multiple codes could be assigned to each review, calculating Cohen's Kappa was not possible. Although there exists no single widely accepted metric for the specific case of multi-label annotation of scientific corpora Ravenscroft et al. [59], for our scenario with up to 17 codes per codebook, the F1 and Boot-F1 metrics by Marchal et al. [47] are applicable. Given that the F1 metric is affected by the prevalence of multi-code assignments by a coder, the Boot-F1 metric corrects this imbalance to make the results comparable across multiple scenarios and pairs of coders.

Table 3 presents the results of the code quality evaluation for the 200 human-coded texts from each of the four topics. In most cases, the highest F1 score is observed between the two human coders, with F1 between C2 and GPT-4o following closely. However, Boot-F1 is consistently highest between C2 and the LLMs, indicating that although the LLMs tend to assign more codes on average than the human coders, they align more closely with C2 than C1 does. This suggests that, when given the same instructions, the LLMs can perform deductive coding with a reliability comparable to a human coder. As GPT-4o outperformed GPT-4 Turbo in the majority of cases across both F1 and Boot-F1, we selected GPT-4o for the full coding.

We noted that the human coders annotated reviews as "Not relevant" more often than the LLM, which may be due to annotation fatigue, where extended coding sessions lead to increased use of default or low-effort labels. On the other hand, it may be that the LLM coding would benefit from explicit few-shot examples [12] of human codes. This could be investigated in future work; here, we abstained from it to avoid overfitting our prompt to our fairly small manually coded test datasets.

5.1.4 Visualizing Trends. To investigate potential temporal changes in the prevalence of the codes, we utilize stacked area charts similar to Experiment 2. For each year, we calculate how often each code category appears, as a percentage of all codes. These percentages are based only on reviews with cosine similarity above a certain threshold. This was done to ensure that the top of chart reproduces the trend curves, and the areas display the relative contribution of codes, while taking into account that more than one code can be assigned to the same review. More precisely, for year y, code c, and experience subscale s, the thickness $\tau(c, y, s)$ of the code's area

	C1 C2		C1 LLM				C2 LLM			
			GPT-4 Turbo		GPT-4o		GPT-4 Turbo		GPT-4o	
Topic	F1	Boot-F1	F1	Boot-F1	F1	Boot-F1	F1	Boot-F1	F1	Boot-F1
Emotional Challenge	**0.574**	0.338	0.420	0.304	0.478	0.353	0.460	0.348	0.545	**0.436**
Boredom	**0.464**	0.392	0.311	0.236	0.396	0.328	0.399	0.326	0.463	**0.400**
Meaning	**0.510**	0.404	0.450	0.361	0.491	0.412	0.479	0.399	0.490	**0.417**
Nostalgia	0.497	0.391	0.469	0.365	0.500	0.400	**0.567**	**0.484**	0.548	0.470

Table 3: Observed F1 and chance-corrected Boot-F1 scores for multi-label inter-coder agreement across the four topics, comparing human coders (C1, C2) and LLM-generated codes from GPT-4 Turbo and GPT-4o. The highest observed and chance-corrected scores are highlighted for each topic, revealing that in some cases, the LLMs achieve greater agreement with C2 than C2 does with C1.

in the plot equals:

$$\tau(c, y, s) = \frac{\text{count}(c, y, s)}{\sum_c \text{count}(c, y, s)} \alpha(y, s). \quad (4)$$

5.2 Experiment 3 Results

Based on the summary of code frequencies in Table 4, the experience trends are produced as the sum of multiple reasons, with no single reason dominating, except for the most frequent reasons for Emotional Challenge, Boredom, and Nostalgia which are almost twice as prevalent as the other reasons. As illustrated in Figure 7, the contributions of each reason have grown steadily over the years, with no clear temporal shifts.

Emotional challenge was most often experienced due to challenging narrative elements (*'This game really shows what it's like to have depression. I've struggled heavily with depression my entire adult life. The text-based choices show the struggle...'*), the game making the player think about real-life issues (*'...This game does well to capture the psychological aspect of one person's personal struggle with domestic violence....'*), and the game making the player feel sad (*'A really emotional and amazing short game.... made me cry at a certain part'*). The coding also revealed other negative emotions, both intentional and unintentional (*'Captivating storyline and challenging enough...Very scary!'*), (*'The Version i played was bugged... Kinda disappointed in that Game.'*).

For Boredom, the most common reason was bad game writing (*'The story is boring as hell and was not interesting at all*'). This was followed by bad sequel/remake/clone games (*'A downgrade from its predecessors...'*) or repetitive/tedious/-grindy gameplay (*'It's weird and consistent, but can get too repetitive and fairly bland after a while. '*).

For Meaning, nostalgia and game stories were the equally most common reasons (*'...glad to play this game again. so much memories and nostalgia within this...'*), (*'This game tells a story worth experiencing and truly shows that video games are an art style...'*).

For nostalgia, the most common reasons were childhood memories of the game or other games (*'I had tons of fun when i played it as a kid, and i think it still holds to this day'*), sequels/prequels (*'It really catches me. It's not as good as the first but still a better sequel than Amnesia: Rebirth'*), and nostalgic/retro graphics/sound/music (*' Definitely worthy game for anyone who is looking to take a hike back to pixelated graphics and cool old-school tunes'*). Childhood memories of something else than a game were also mentioned, but this was much less frequent and the memories also involved a game (*'This game came out with nothing, but the friends I had at the time to play with made this game everything it could be. Some of the funniest times I ever had were on this game...'*).

6 Experiment 4: Correlations of Experience Constructs and Review Scores

Now that we can estimate aspects of player experience from existing review data, a crucial question emerges: Which aspects of player experience are the most relevant to a game's success? What should game developers prioritize, with their inevitably limited resources? To probe these questions, we calculated the Pearson correlations between the review scores and the review-subscale similarities of Equation (2), i.e., the estimated agreement of the reviews with the items of each subscale. The correlations were computed using the Python Pingouin package, controlling for review year and the platform of the reviewed game.

The results are presented in Table 5, sorted from highest to lowest correlation for each questionnaire. Positive correlations indicate experiences associated with high review scores (e.g., Joy, Enchantment, Mastery, Audiovisual Appeal), and negative correlations indicate experiences associated with low review scores (e.g., Feeling of Ugliness, Boredom). Due to our large sample size, even the weak correlations are statistically significant ($p<0.001$).

7 Discussion

Considering the results of all our experiments together, we can now synthesize and discuss the main findings of our study. A general conclusion that one can draw based on all the trend breakdowns into genres, games, and qualitative codes is that the trends observable in our data are produced by multiple compounding factors, with no clear dominating contributors such as a single landmark game or genre. The notable exceptions are the relatively higher prevalence of bad game writing as a reason for Boredom, childhood memories of games as a reason for Nostalgia, and challenging narrative elements as

Emotional Challenge	Boredom	Meaning	Nostalgia
Challenging narrative elements (24.9%)	Bad/uninteresting story or dialogue (35.6%)	Nostalgia (24.8%)	Childhood memories of the game or other games (24.5%)
The game made me think about real-life issues (14.9%)	Bad sequel, remake, or clone of some other game (17.7%)	Story (24.6%)	Sequel/prequel (14.8%)
The things that happened in the game made me sad (11.9%)	Repetitive/tedious/grindy (16.5%)	Emotional impact (20.4%)	Nostalgic/retro graphics/sound/music (14.1%)
I felt a sense of responsibility for characters and events in the game (11.3%)	Lack of depth (14.9%)	Unique game/pushes the boundaries/made me realise something about games (20.2%)	Rediscovering an old game/franchise or coming back to gaming after a break (9.1%)
The game had moral dilemmas in it where the choice was not obvious (9.8%)	Bad/uninteresting audiovisual (14.5%)	Memorable (13.5%)	Remake/recreation/mod (8.2%)
Other negative emotions, intentional (9.2%)	Bad/uninteresting level design (14.5%)	Other personal impact (9.6%)	First game/genre I played (6.7%)
Other negative emotions, unintentional (6.1%)	Good start turned boring (13.6%)	Overcame a difficult challenge (7.3%)	Big personal impact (3.8%)
The game involved making moral choices that I didn't agree with (5.4%)	Lack of agency/choice (10.0%)	Thought-provoking (6.3%)	Childhood memories of something else than a game (2.4%)
	Unoriginal (10.0%)	Historical/cultural immersion or references (6.1%)	
	Lack of challenge/too easy (8.9%)	Motivational impact (4.4%)	
	Slow pacing (8.5%)	Learned something from the game (2.3%)	
	Low replayability (5.1%)		
	Tech/performance issues (3.8%)		
	Lack of feedback (1.7%)		
Not relevant (40.5%)	Not relevant (8.7%)	Not relevant (17.6%)	Not relevant (30.0%)
No reason (10.3%)	No reason (18.8%)	No reason (8.9%)	No reason (13.0%)
Other reason (4.4%)	Other reason (3.0%)	Other reason (9.2%)	Other reason (5.7%)

Table 4: The reasons for experiencing Emotional Challenge, Boredom, Meaning and Nostalgia, sorted according to frequencies.

a reason for Emotional Challenge. Longitudinally, the various contributors to the trends have grown steadily, except for the Point-and-Click genre dominating Emotional Challenge in 2012-2016 but then declining, and the Survival genre emerging as a contributor to Emotional Challenge in 2019.

Below, we discuss the results from the point of view of player experiences evolving both for the better and worse, and what actionable insights our study can provide for game designers and developers. Finally, we discuss some observations our work provides about the questionnaires used as well as the limitations of our work.

7.1 Maturation of Games as an Art Form

Notably, our findings revealed an increasing trend of game reviews describing emotionally challenging, meaningful, and nostalgic experiences, which we interpret as a welcome sign of the maturation of games as an art form.

Our results echo how entertainment research—particularly related to the medium of games—has been increasingly interested in the idea of *eudaimonia* [24, 25, 57]. While player experience research has traditionally focused on purely positive emotions, such as fun and enjoyment, the past years have seen increased attention to how also negative, or mixed emotions can lead to positive player experiences, e.g., with players appreciating their emotionally moving or sad gameplay encounters [7]. In the broader context of entertainment media, gratifications associated with the enjoyment of positive, inherently pleasurable emotions are often characterized as *hedonistic*, whereas those related to appreciating the moving, meaningful, or thought-provoking aspects of an experience are characterized as *eudaimonic* [51]. This dichotomy between hedonia and eudaimonia—both concepts dating back to ancient Greece [26]—has also recently gained increased interest

Figure 7: The contributions of the coded reasons to experience trends. There are no clear temporal changes, the prevalence of all the reasons growing somewhat steadily over time.

within HCI [48], particularly when it comes to research on games [10, 21, 24, 25, 57, 70].

Meaningfulness, emotional challenge, and nostalgia are all concepts that game researchers have related to eudaimonic gaming experiences [24, 57]. Hence, our findings give credence to the increasing academic interest towards eudaimonic gaming and imply that there might be an upward trend in such player experiences, or, at least in players' readiness to discuss such experiences in game reviews. To yield further insights into eudaimonic player experiences—and potential trends related to it—we suggest that future research could expand our analyses to focus on a more comprehensive set of eudaimonia-related concepts (see [25, p.5] for an overview), including, perhaps, concepts like reflection, personal growth, connectedness, appreciation, and eudaimonic/psychological well-being.

According to our data, nostalgic game experiences are associated with high review scores—albeit weakly—and also contribute to games being experienced as meaningful. The most common reasons the reviews express for nostalgia are what one would expect: Childhood memories of the game or other games, sequels/prequels/remakes/mods, nostalgic/retro graphics/sound/music, or rediscovering and old game or franchise, among other reasons. Only in rare cases was nostalgia evoked by childhood memories of something else than a game. These findings are echoed by Bowman and Wulf [11], who, in their recent study on nostalgia in video games, highlighted how the average gamer has been estimated to be in their 30s—oftentimes playing together with their children—while video games are becoming more and more ubiquitous and mature as a medium [11]. Moreover, there has been a growth of so-called 'retro games' and the gaming industry has seen a trend in developing video games with a specific emphasis on nostalgia [11]. In fact, the phenomenon of nostalgia in retro game design has been argued to corroborate the position of video games as a mature art form capable of evoking reflective longing [33].

While the experience of nostalgia in games has been connected to eudaimonia [24, 25], within HCI, nostalgia seems to have gained less academic attention than other eudaimonia-related concepts [56], such as notions of reflective, meaningful, emotionally challenging, or transformative experiences [e.g., 18, 20, 22, 39, 49, 58, 70, 74, 75]. In entertainment research more generally, media-induced nostalgia—a self-conscious, bittersweet, fundamentally social, and predominantly positive emotion that stems from past memories and yearning [65]—has been seen to serve self-oriented, existential, and social functions, contributing both to hedonic and eudaimonic entertainment experiences and supporting well-being [79]. As with other media, nostalgia induced by videogames has been theorized and shown to positively influence players subjective

Questionnaire	Subscale	*r*	p
PXI	Mastery	0.40	0.000***
	Audiovisual Appeal	0.39	0.000g
	Autonomy	0.32	0.000***
	Meaning	0.31	0.000***
	Immersion	0.24	0.000***
	Challenge	0.23	0.000***
	Curiosity	0.21	0.000***
	Progress Feedback	0.20	0.000***
	Goals and Rules	0.17	0.000***
	Ease of Control	0.10	0.000***
CORGIS	Performative Challenge	0.18	0.000***
	Cognitive Challenge	0.14	0.000***
	Emotional Challenge	0.09	0.000***
	Decision Making Challenge	0.02	0.000***
AESTHEMOS	Joy	0.53	0.000***
	Enchantment	0.50	0.000***
	Feeling of Beauty/Liking	0.49	0.000***
	Fascination	0.48	0.000***
	Relaxation	0.48	0.000***
	Awe	0.46	0.000***
	Vitality	0.42	0.000***
	Energy	0.42	0.000***
	Nostalgia	0.37	0.000***
	Being moved	0.36	0.000***
	Interest	0.35	0.000***
	Intellectual Challenge	0.34	0.000***
	Insight	0.33	0.000***
	Humor	0.25	0.000***
	Surprise	0.19	0.000***
	Sadness	-0.11	0.000***
	Confusion	-0.28	0.000***
	Uneasiness	-0.28	0.000***
	Anger	-0.36	0.000***
	Boredom	-0.40	0.000***
	Feeling of Ugliness	-0.58	0.000***

Table 5: Correlations (*r*) between the review scores and the estimated review-subscale agreements, highest correlations first. A high correlation suggests a high priority for a game that aims for good reviews. Due to our large sample size, even the small correlations are statistically significant (p<0.001 denoted by *).**

well-being (e.g., self-reported life satisfaction) as well as psychological well-being (e.g., personal flourishing) [11]. Some such effects of nostalgia on players' entertainment experiences and well-being have been studied in the context of *Pokémon Go* [78–80]. Our findings suggest that there has been a general rise in players' experiences of nostalgia in games, and we propose that further tapping into online data sources where players talk about nostalgia could pose a fruitful avenue for future research.

7.2 Boredom on the Rise

On the negative side, it appears that players are increasingly experiencing boredom, which is also associated with low review scores. The most frequent reasons the reviews indicate for boredom are bad writing, bad sequels/remakes/clone games, and repetitive/tedious/grindy gameplay. This might reflect the democratization of game development tools and distribution channels; lowering the barrier of entry inevitably floods the market with low-quality games and may make it harder to discover the higher quality games. The trend might also reflect

that as more and more games get made and the tools and technologies of game creation mature, it is perhaps increasingly difficult to invent truly novel and original games.

On the other hand, one should keep in mind that our observed association of boredom and low review scores only holds on average, and there can also be games and contexts where some players find meaning in boredom. For instance, prolonged moments of boredom which require the player to slow down may contribute to memorable experiences of novelty or uniqueness which stand out from regular gameplay moments [68].

Considering the future, it will be interesting to observe how the proliferation of AI will impact the amount of boring games. AI appears to be taking the democratization and empowering of game creation even further, to its logical extreme: Platforms such as https://bitmagic.ai/ and https://rosebud.ai/ are already promising that one can create one's dream game simply by describing it in natural language and/or with back-of-the-napkin sketches. The pessimistic outcome is that unoriginal and boring "AI slop" games will flood the market and make it even harder to discover high-quality content. Then again, the overwhelming quantity of AI-generated games might result in them having zero perceived value to others than their creators. Therefore, AI-generated games might primarily exist as a form of AI-assisted self-expression on dedicated user-generated content platforms instead of commercial game marketplaces such as Steam, especially if the marketplaces improve their moderation tools and processes to disincentivise AI slop. In the right hands, AI may also be beneficial, as it holds the promise of enabling more ambitious games with smaller teams and budgets.

7.3 Insights for Game Developers

We consider our results to provide two primary benefits for game designers and developers. First, we provide a new way to discover reference games for designing particular types of player experiences. We demonstrate this in Table 2, where clicking on a game reference such as [G32] takes one to the ludography, which then provides links to more information such as the game's Steam store page with a description, videos, and screenshots.

Second, our results highlight that if one's goal is to achieve high reviews, all aspects of player experience are not equal. Here, our correlation data in Table 5 may help designers in determining what to prioritize, given their inevitably limited resources.

For instance, although it might be tempting to additively combine the PXI subscale scores to an overall player experience score, our correlation data suggests that prioritizing Mastery, Audiovisual Appeal, Autonomy, and Meaning may provide a higher return on investment than Ease of Control and (the clarity of) Goals and Rules. For instance, over-prioritizing the latter might result in too heavy-handed tutorialization or a too simplistic design. Perhaps this means that more developers should follow in the footsteps of games like Elden Ring [G18], a recent extremely successful and complex game with almost no tutorialization.

Similarly, not all challenge types appear equally important. There appears to be no correlation between review scores and decision making challenge, and among the rest of the CORGIS challenge types, the correlation for emotional challenge is weakest. However, as elaborated below, this may partially be due to the way CORGIS operationalizes emotional challenge.

7.4 How Should One Measure Emotional Challenge?

The correlations of Table 5 suggest that emotional challenge is only weakly associated with high review scores, even though players have been found to appreciate emotionally challenging games [8]. This prompts further investigation, for which we provide some initial remarks below.

In Experiment 3, both our data coders found the concept of emotional challenge hard to understand, and asked us to clarify whether something is emotionally challenging if it agrees with all the CORGIS subscale items, on average, or at least one or some of them. We settled for the average, as when analyzing human responses, the questionnaire items are combined additively. The data coding also resulted in a high percentage of reviews (40%) that the coders flagged as not relevant to emotional challenge, suggesting that the thresholding based on embedding similarity was not very accurate in detecting relevant reviews.

One possible reason for the issues is that the emotional challenge items might not measure the construct as coherently as the other CORGIS subscales. To investigate the weak correlation with review score, we provide the individual item-level correlations in Appendix B. These exhibit considerable spread, with some emotional challenge items associated with high review scores, and others associated with low scores, which explains the overall weak correlation. Regarding alternative ways to operationalize emotional challenge, Flint et al. [32] have explored combining CORGIS items with those of the Video Game Demand Scale (VGDS). They discovered a seven-factor solution where the emotional challenge/demand category included only one CORGIS emotional challenge item ("The things that happened in the game made me sad") and 5 VGDS items (e.g., "I had a lot of unexpected feelings during gameplay" and "I was moved by the game".).

8 Limitations and Future Work

An obvious limitation of our work is that while game reviews no doubt reflect player experiences, to some degree, we can only directly measure trends in review discourse instead of actual experience trends. Confounding factors such as language drift and shifting evaluation norms can also influence the share of reviews above threshold. Moreover, as the Metacritic review data does not include ground truth responses to any questionnaire, we cannot quantify how accurately the embedding-based semantic similarity maps to each particular reviewer's agreement with questionnaire items. However, our

validation in Appendix A does indicate that the semantic similarity is predictive of review-item agreement rated by human coders, which should be correlated with the reviewer's actual agreement, assuming that the reviews are honest and without excessive sarcasm that our human coders were not able to detect. To mitigate these issues, future work should aim to collect a ground truth dataset of game reviews combined with questionnaire responses. Such data could then be used to evaluate and improve the accuracy of both embedding-based and other approaches to predicting the responses based on the reviews.

Considering that the population of players writing reviews has no doubt evolved over the years, measurement invariance is an additional potential error source. So far, there are no studies about the measurement invariance of AESTHEMOS, PXI, or CORGIS across different populations or time. Generally, measurement invariance is hard to achieve [71] and studying it appears rare in player experience literature. The BANGS scale provides a rare exception where the validation does also include measurement invariance [6].

The brevity and variable quality of the user-written reviews poses further limitations. Future work could replicate our analyses with expert reviews that are more likely to discuss games more thoughtfully, from multiple angles. However, calculating a single embedding vector based on long expert review might not be informative, and depending on the number of reviews to analyze, it might be better to adopt a prompting-based approach (see Section 3.2.3) or investigate how to best divide a review into multiple chunks that are embedded independently.

Finally, due to space and resource limitations, we could not visualize and analyze all trends with the same detail and had to apply some difficult prioritization. However, our approach should be straightforward for others to implement, building on our Python source code. Future work might look into applying our approach using different questionnaires such as the BANGS scale mentioned above, as well as different data such as social media discussions. At the same time, review data which is naturally paired with games or other products does provide extra opportunities such as using the item-review similarities to discover games that score particularly high or low on specific experience dimensions. Table 2 provides an example of this in case of high-scoring games. Considering low-scoring games, the scatterplot in Appendix A suggests that cosine similarity may be less reliable in measuring disagreement than agreement. Therefore, to identify low-scoring games, one might have to re-word scale items as their negative counterparts and then find the high-scoring games.

9 Conclusion

We have presented a novel and highly scalable approach for mapping game reviews to player experience questionnaire items. The approach is demonstrated with a large dataset of Metacritic reviews and the PXI, CORGIS, and AESTHEMOS questionnaires. Our investigation highlights a rising trend of eudaimonic game experiences (e.g., emotional challenge, meaning, nostalgia) fueled by aspects such as emotional narratives and games dealing with real-life issues. We interpret this as a welcome sign of the maturation of games as an artform. On the negative side, we also observe a trend of increasing boredom caused by bad game writing and bad sequels, remakes, and clone games among other reasons. For game designers and developers, our work provides two main benefits. First, our correlation data of experiences and review scores may be helpful in prioritizing resources. Second, we demonstrate the usability of the embedding similarity of reviews and questionnaire items in discovering reference games for designing particular types of experiences.

Acknowledgments

This research has been funded by Finland's Ministry of Education and Culture's Doctoral Education Pilot under Decision No. VN/3137/2024-OKM-6 (The Finnish Doctoral Program Network in Artificial Intelligence, AI-DOC) and Research Council of Finland (MAGE – Establishing a Model of Aesthetic Game Experience, 339350).

References

[1] Vero Vanden Abeele, Katta Spiel, Lennart Nacke, Daniel Johnson, and Kathrin Gerling. 2020. Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies* 135 (2020), 102370.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report.

[3] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The Illusion of Artificial Inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 677, 21 pages. doi:10.1145/3613904.3642703

[4] Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. 2023. Using Large Language Models for Qualitative Analysis can Introduce Serious Bias. *arXiv preprint arXiv:2309.17147* abs/2309.17147 (2023), 1–37. doi:10.48550/ARXIV.2309.17147

[5] Mohammad Atari, Ali Omrani, and Morteza Dehghani. 2023. Contextualized construct representation: leveraging psychometric scales to advance theory-driven text analysis. *PsyArXiv preprint* m93pd (2023), 1–32. doi:10.31234/osf.io/m93pd

[6] Nick Ballou, Alena Denisova, Richard Ryan, C Scott Rigby, and Sebastian Deterding. 2024. The Basic Needs in Games Scale (BANGS): A new tool for investigating positive and negative video game experiences. *International Journal of Human-Computer Studies* 188 (2024), 103289.

[7] Julia Ayumi Bopp, Elisa D. Mekler, and Klaus Opwis. 2016. Negative emotion, positive experience? Emotionally moving moments in digital games. *Conference on Human Factors in Computing Systems - Proceedings* 2016-May (2016), 2996–3006. doi:10.1145/2858036.2858227 ISBN: 9781450333627.

[8] Julia Ayumi Bopp, Klaus Opwis, and Elisa D. Mekler. 2018. "An odd kind of pleasure": Differentiating emotional challenge in digital games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173615

[9] Julia A Bopp, Jan B Vornhagen, and Elisa D Mekler. 2021. " My Soul Got a Little Bit Cleaner" Art Experience in Videogames. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 19.

[10] Nicholas David Bowman, Rowan Daneels, and Daniel Possler. 2024. Excited for eudaimonia? An emergent thematic analysis of player expectations of upcoming video games. *Psychology of Popular Media* 13, 3 (2024), 416–427. doi:10.1037/ppm0000474

[11] Nicholas David Bowman and Tim Wulf. 2023. Nostalgia in video games. *Current Opinion in Psychology* 49 (2023), 101544. doi:10.1016/j.copsyc.2022.101544

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry,

Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[13] Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. 10853–10875 pages.

[14] Gifford K Cheung, Thomas Zimmermann, and Nachiappan Nagappan. 2014. The first hour experience: how the initial play can engage (or lose) new players. 57–66 pages.

[15] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-assisted content analysis: Using large language models to support deductive coding.

[16] Tom Cole, Paul Cairns, and Marco Gillies. 2015. Emotional and functional challenge in core and avant-garde games. 121–126 pages.

[17] Tom Cole, Paul Cairns, and Marco Gillies. 2015. Emotional and functional challenge in core and avant-garde games. 121–126 pages.

[18] Tom Cole, Paul Cairns, and Marco Gillies. 2015. Emotional and functional challenge in core and avant-garde games. 121–126 pages. doi:10.1145/2793107.2793147 ISBN: 9781450334662.

[19] Tom Cole and Marco Gillies. 2021. Thinking and doing: Challenge, agency, and the eudaimonic experience in video games. *Games and Culture* 16, 2 (2021), 187–207.

[20] Tom Cole and Marco Gillies. 2021. Thinking and Doing: Challenge, Agency, and the Eudaimonic Experience in Video Games. *Games and Culture* 16, 2 (2021), 187–207. doi:10.1177/1555412019881536

[21] Tom Cole and Marco Gillies. 2022. Emotional exploration and the eudaimonic gameplay experience: a grounded theory. 16 pages.

[22] Tom Cole and Marco Gillies. 2022. Emotional Exploration and the Eudaimonic Gameplay Experience: A Grounded Theory. doi:10.1145/3491102.3502002 event-place: New Orleans, LA, USA.

[23] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9993–10001. doi:10.18653/v1/2023.findings-emnlp.669

[24] Rowan Daneels, Nicholas D Bowman, Daniel Possler, and Elisa D Mekler. 2021. The 'eudaimonic experience': A scoping review of the concept in digital games research. *Media and Communication* 9, 2 (2021), 178–190.

[25] Rowan Daneels, Heidi Vandebosch, and Michel Walrave. 2023. " Deeper gaming": a literature review and research agenda on eudaimonia in digital games research. *Technology, mind, and behavior* 4, 2 (2023), 1–13.

[26] Edward L. Deci and Richard M. Ryan. 2008. Hedonia, eudaimonia, and well-being: An introduction. *Journal of Happiness Studies* 9, 1 (2008), 1–11. doi:10.1007/s10902-006-9018-1

[27] Fatemeh Dehghani and Loutfouz Zaman. 2024. Exploring Players' Perspectives: A Comprehensive Topic Modeling Case Study on Elden Ring. *Information* 15, 9 (2024), 573.

[28] Alena Denisova, Paul Cairns, Christian Guckelsberger, and David Zendle. 2020. Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS). *International Journal of Human-Computer Studies* 137 (2020), 102383.

[29] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models.

[30] Zackary Okun Dunivin. 2024. Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks. doi:10.48550/ARXIV.2401.15170

[31] Hugging Face. 2025. MTEB Leaderboard. https://huggingface.co/spaces/mteb/leaderboard. Accessed: 2025-02-04.

[32] Alex Flint, Alena Denisova, and Nick Bowman. 2023. Comparing Measures of perceived challenge and demand in video games: Exploring the conceptual dimensions of CORGIS and VGDS. 19 pages.

[33] Maria B Garda. 2013. Nostalgia in retro game design.

[34] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. 19 pages.

[35] Leah Hamilton, Desha Elliott, Aaron Quick, Simone Smith, and Victoria Choplin. 2023. Exploring the Use of AI in Qualitative Analysis: A Comparative Study of Guaranteed Income Data. *International Journal of Qualitative Methods* 22 (Oct. 2023), 16094069231201504. doi:10.1177/16094069231201504

[36] Jimpei Hitsuwari, Hirohito Okano, and Michio Nomura. 2024. Predicting attitudes toward ambiguity using natural language processing on free descriptions for open-ended question measurements. *Scientific Reports* 14, 1 (2024), 8276.

[37] Feng Huang, Xia Sun, Aizhu Mei, Yilin Wang, Huimin Ding, and Tingshao Zhu. 2024. LLM Plus Machine Learning Outperform Expert Rating to Predict Life Satisfaction From Self-Statement Text.

[38] Player Experience Inventory. 2025. Player Experience Inventory Instrument. https://playerexperienceinventory.org/instrument. Accessed: 2025-02-04.

[39] Rilla Khaled. 2018. Questions Over Answers: Reflective Game Design. 3–27 pages. doi:10.1007/978-981-10-1891-6_1

[40] J Matias Kivikangas, Guillaume Chanel, Ben Cowley, Inger Ekman, Mikko Salminen, Simo Järvelä, and Niklas Ravaja. 2011. A review of the use of psychophysiological methods in game research. *journal of gaming & virtual worlds* 3, 3 (2011), 181–199.

[41] Oscar NE Kjell, Katarina Kjell, Danilo Garcia, and Sverker Sikström. 2019. Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods* 24, 1 (2019), 92.

[42] Oscar NE Kjell, Katarina Kjell, and H Andrew Schwartz. 2024. Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research* 333 (2024), 115667.

[43] Oscar NE Kjell, Sverker Sikström, Katarina Kjell, and H Andrew Schwartz. 2022. Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific reports* 12, 1 (2022), 3918.

[44] Aris Kosmopoulos, Antonios Liapis, George Giannakopoulos, and Nikiforos Pittaras. 2020. Summarizing Game Reviews: First Contact. 22–31 pages.

[45] Michael Lankes and Andreas Stöckl. 2023. Game reviews reviewed: A game designer's perspective on AI-generated game review analyses. 8 pages.

[46] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (Beijing, China) *(ICML'14)*. JMLR.org, United States, II–1188–II–1196.

[47] Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing Annotation Quality in Multi-label Annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 3659–3668. https://aclanthology.org/2022.coling-1.322/

[48] Elisa D. Mekler and Kasper Hornbæk. 2016. Momentary Pleasure or Lasting Meaning? Distinguishing Eudaimonic and Hedonic User Experiences.

[49] Elisa D. Mekler, Ioanna Iacovides, and Julia Ayumi Bopp. 2018. "A Game that Makes You Question...": Exploring the Role of Reflection for the Player Experience. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 315–327. doi:10.1145/3242671.3242691

[50] Matthew Miles, A. Michael Huberman, and Johnny Saldaña. 2014. Qualitative Data Analysis: A Methods Sourcebook.

[51] Mary Beth Oliver and Anne Bartsch. 2010. Appreciation as Audience Response: Exploring Entertainment Gratifications Beyond Hedonism. *Human Communication Research* 36, 1 (Jan. 2010), 53–81. doi:10.1111/j.1468-2958.2009.01368.x

[52] OpenAI. 2024. OpenAI API Documentation: Embeddings. https://platform.openai.com/docs/guides/embeddings

[53] Sicheng Pan, Gary JW Xu, Kun Guo, Seop Hyeong Park, and Hongliang Ding. 2024. Cultural insights in souls-like games: analyzing player behaviors, perspectives, and emotions across a multicultural context.

[54] Heinrich Peters and Sandra C Matz. 2024. Large language models can infer psychological dispositions of social media users. *PNAS nexus* 3, 6 (2024), pgae231.

[55] Cody Phillips, Madison Klarkowski, Julian Frommel, Carl Gutwin, and Regan L Mandryk. 2021. Identifying commercial games with therapeutic potential through a content analysis of steam reviews. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 1–21.

[56] Daniel Possler, Nicholas David Bowman, and Rowan Daneels. 2023. Explaining the formation of eudaimonic gaming experiences: a theoretical overview and systemization based on interactivity and game elements. *Frontiers in Communication* 8 (Sept. 2023), 1215960. doi:10.3389/fcomm.2023.1215960

[57] Daniel Possler, Rowan Daneels, and Nicholas D. Bowman. 2024. Players Just Want to Have Fun? An Exploratory Survey on Hedonic and Eudaimonic Game Motives. *Games and Culture* 19, 5 (July 2024), 611–633. doi:10.1177/15554120231182498

[58] Heidi Rautalahti. 2019. "How video games changed my life": Life-Changing Testimonies and The Last of Us. *gamevironments* 10 (2019), 1–38. http://www.gamevironments.uni-bremen.de/

[59] James Ravenscroft, Anika Oellrich, Shyamasree Saha, and Maria Liakata. 2016. Multi-label Annotation in Scientific Articles - The Multi-label Cancer Risk Assessment Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and

Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Portorož, Slovenia, 4115–4123. https://aclanthology.org/L16-1650/

[60] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

[61] Shaghayegh Roohi, Jari Takatalo, J Matias Kivikangas, and Perttu Hämäläinen. 2018. Neural network based facial expression analysis of gameevents: a cautionary tale. 429–437 pages.

[62] Gony Rosenman, Lior Wolf, and Talma Hendler. 2024. LLM Questionnaire Completion for Automatic Psychiatric Assessment.

[63] Tiago Santos, Florian Lemmerich, Markus Strohmaier, and Denis Helic. 2019. What's in a review: Discrepancies between expert and amateur reviews of video games on metacritic. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 22.

[64] Ines Schindler, Georg Hosoya, Winfried Menninghaus, Ursula Beermann, Valentin Wagner, Michael Eid, and Klaus R Scherer. 2017. Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PloS one* 12, 6 (2017), e0178899.

[65] Constantine Sedikides, Tim Wildschut, Clay Routledge, Jamie Arndt, Erica G. Hepper, and Xinyue Zhou. 2015. To Nostalgize: Mixing Memory with Affect and Desire. In *Advances in Experimental Social Psychology*. Vol. 51. Academic Press (Elsevier), San Diego, CA, USA, 189–273. doi:10.1016/bs.aesp.2014.10.001

[66] Ravi Sinha, Idris Solola, Ha Nguyen, Hillary Swanson, and LuEttaMae Lawrence. 2024. The Role of Generative AI in Qualitative Research: GPT-4's Contributions to a Grounded Theory Analysis. In *Proceedings of the 3rd International Conference on Learning, Design, and Technology* (Evanston, IL, USA) *(LDT '24)*. Association for Computing Machinery, New York, NY, USA, 17–25. doi:10.1145/3663433.3663456

[67] Mikke Tavast, Anton Kunnari, and Perttu Hämäläinen. 2022. Language models can generate human-like self-reports of emotion. 69–72 pages.

[68] Nina Tepponen, Prabhav Bhatnagar, Jaakko Väkevä, and Perttu Hämäläinen. 2025. Towards Understanding Waiting in Video Games.

[69] April Tyack and Elisa D Mekler. 2020. Self-determination theory in HCI games research: Current uses and open questions. 22 pages.

[70] Jaakko Väkevä, Elisa D. Mekler, and Janne Lindqvist. 2024. From Disorientation to Harmony: Autoethnographic Insights into Transformative Videogame Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 808, 20 pages. doi:10.1145/3613904.3642543

[71] Rens Van De Schoot, Peter Schmidt, Alain De Beuckelaer, Kimberley Lek, and Marielle Zondervan-Zwijnenburg. 2015. Measurement invariance. 1064 pages.

[72] Xiaohui Wang and Dion Hoe-Lian Goh. 2020. Components of game experience: An automatic text analysis of online reviews. *Entertainment Computing* 33 (2020), 100338.

[73] Zhiqiang Wang, Yiran Pang, Yanbin Lin, and Xingquan Zhu. 2024. Adaptable and reliable text classification using large language models. 67–74 pages.

[74] Matthew Alexander Whitby, Sebastian Deterding, and Ioanna Iacovides. 2019. "One of the baddies all along": Moments that challenge a player's perspective. 339–350 pages. doi:10.1145/3311350.3347192 ISBN: 9781450366885.

[75] Matthew Alexander Whitby, Ioanna Iacovides, and Sebastian Deterding. 2023. "Conversations with pigeons": Capturing Players' Lived Experience of Perspective Challenging Games. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (Sept. 2023), 833–855. doi:10.1145/3611051

[76] Josef Wiemeyer, Lennart Nacke, Christiane Moser, and Florian 'Floyd'Mueller. 2016. Player experience. 243–271 pages.

[77] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.

[78] Tim Wulf and Matthew Baldwin. 2020. Being a kid again: Playing Pokémon Go contributes to wellbeing through nostalgia. *Studies in Communication and Media* 9, 2 (2020), 241–263. doi:10.5771/2192-4007-2020-2-241

[79] Tim Wulf, Diana Rieger, and Josephine B. Schmitt. 2018. Blissed by the past: Theorizing media-induced nostalgia as an audience response factor for entertainment and well-being. *Poetics* (To be filled if known) (2018), (To be filled if known). doi:10.1016/j.poetic.2018.04.001

[80] Chia-chen Yang and Dong Liu. 2017. Motives Matter: Motives for Playing Pokémon Go and Implications for Well-Being. *Cyberpsychology, Behavior, and Social Networking* 20, 1 (Jan. 2017), 52–57. doi:10.1089/cyber.2016.0562

[81] Yang Yu, Tai Dinh, Fangyu Yu, and Van-Nam Huynh. 2023. Understanding Mobile Game Reviews Through Sentiment Analysis: A Case Study of PUBGm. 102–115 pages.

[82] Yang Yu, Ba-Hung Nguyen, Fangyu Yu, and Van-Nam Huynh. 2021. Discovering topics of interest on Steam community using an LDA approach. 510–517 pages.

[83] José P Zagal, Noriko Tomuro, and Andriy Shepitsen. 2012. Natural language processing in game studies research: An overview. *Simulation & Gaming* 43, 3 (2012), 356–373.

[84] Miaoqi Zhu and Xiaowen Fang. 2015. A lexical approach to study computer games and game play experience via online reviews. *International Journal of Human-Computer Interaction* 31, 6 (2015), 413–426.

[85] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291.

Ludography

[G1] Awaceb . 2024. Tchia on Steam. Digital Game. https://store.steampowered.com/app/1496590/Tchia/

[G2] BADGAMES. 2022. drowning on Steam. Digital Game. https://store.steampowered.com/app/1890340/drowning/

[G3] BANDAI NAMCO Studio Inc. . 2015. Tales of Zestiria on Steam. Digital Game. https://store.steampowered.com/app/351970/Tales_of_Zestiria/

[G4] Beethoven and Dinosaur . 2021. The Artful Escape on Steam. Digital Game. https://store.steampowered.com/app/1122680/The_Artful_Escape/

[G5] Ben Esposito . 2018. Donut County on Steam. Digital Game. https://store.steampowered.com/app/702670/Donut_County/

[G6] Campo Santo . 2016. Firewatch on Steam. Digital Game. https://store.steampowered.com/app/383870/Firewatch/

[G7] Cococucumber . 2023. Ravenlok | Download and Buy Today - Epic Games Store. Digital Game. https://store.epicgames.com/en-US/p/ravenlok-bcabbc

[G8] Coldwood Interactive . 2016. Unravel on Steam. Digital Game. https://store.steampowered.com/app/1225560/Unravel/

[G9] D-Pad Studio . 2016. Owlboy on Steam. Digital Game. https://store.steampowered.com/app/115800/Owlboy/

[G10] DICE . 2016. Mirror's Edgeâ„¢ Catalyst on Steam. Digital Game. https://store.steampowered.com/app/1233570/Mirrors_Edge_Catalyst/

[G11] Digixart . 2021. Road 96 on Steam. Digital Game. https://store.steampowered.com/app/1466640/Road_96/

[G12] Dingaling . 2014. LISA Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/lisa/

[G13] DONTNOD Entertainment . 2015. Life is Strange Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/life-is-strange/

[G14] DONTNOD Entertainment . 2020. Tell Me Why on Steam. Digital Game. https://store.steampowered.com/app/1180660/Tell_Me_Why/

[G15] Everything Unlimited Ltd. . 2015. The Beginner's Guide on Steam. Digital Game. https://store.steampowered.com/app/303210/The_Beginners_Guide/

[G16] Fallen Leaf, Black Drakkar Games . 2023. Fort Solis on Steam. Digital Game. https://store.steampowered.com/app/1931730/Fort_Solis/

[G17] Freebird Games . 2011. To the Moon on Steam. Digital Game. https://store.steampowered.com/app/206440/To_the_Moon/

[G18] FromSoftware, Inc. . 2022. Elden Ring. Digital Game. https://store.steampowered.com/app/1245620/ELDEN_RING/

[G19] Fullbright . 2013. Gone Home on Steam. Digital Game. https://store.steampowered.com/app/232430/Gone_Home/

[G20] Galactic Cafe . 2022. The Stanley Parable: Ultra Deluxe on Steam. Digital Game. https://store.steampowered.com/app/1703340/The_Stanley_Parable_Ultra_Deluxe/

[G21] Geometric Interactive . 2023. COCOON on Steam. Digital Game. https://store.steampowered.com/app/1497440/COCOON/

[G22] Giant Squid . 2016. ABZU on Steam. Digital Game. https://store.steampowered.com/app/384190/ABZU/

[G23] Giant Squid. 2021. The Pathless on Steam. Digital Game. https://store.steampowered.com/app/1492680/The_Pathless/

[G24] Going Loud Studios. 2013. DLC Quest on Steam. Digital Game. https://store.steampowered.com/app/230050/DLC_Quest/

[G25] Good-Feel . 2024. Princess Peachâ„¢: Showtime! for Nintendo Switch - Nintendo Official Site. Digital Game. https://www.nintendo.com/us/store/products/princess-peach-showtime-switch/?srsltid=AfmBOoq6cP4uqTk12YDhWOXTafvYLms5u92x5Dg5hf_yq5K7AbFvE_Lm

[G26] Handheld Games . 2008. Disney Cory in the House Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/disney-cory-in-the-house/

[G27] Harvester Games . 2012. The Cat Lady on Steam. Digital Game. https://store.steampowered.com/app/253110/The_Cat_Lady/

[G28] High Moon Studios . 2013. Deadpool Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/deadpool/

[G29] Housemarque, Climax Studios . 2023. Returnalâ„¢ on Steam. Digital Game. https://store.steampowered.com/app/1649240/Returnal/
[G30] Kyle Seeley . 2015. Emily is Away on Steam. Digital Game. https://store.steampowered.com/app/417860/Emily_is_Away/
[G31] Lovable Hat Cult . 2020. Journey of the Broken Circle on Steam. Digital Game. https://store.steampowered.com/app/1179620/Journey_of_the_Broken_Circle/
[G32] MidBoss, LLC. . 2015. 2064: Read Only Memories on Steam. Digital Game. https://store.steampowered.com/app/330820/2064_Read_Only_Memories/
[G33] Moon Studios GmbH . 2016. Ori and the Blind Forest: Definitive Edition on Steam. Digital Game. https://store.steampowered.com/app/387290/Ori_and_the_Blind_Forest_Definitive_Edition/
[G34] Necrophone Games . 2014. Jazzpunk: Director's Cut on Steam. Digital Game. https://store.steampowered.com/app/250260/Jazzpunk_Directors_Cut/
[G35] Nightdive Studios . 2023. System Shock on Steam. Digital Game. https://store.steampowered.com/app/482400/System_Shock/
[G36] Nintendo. 2023. Super Mario Bros. Wonder Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/super-mario-bros-wonder/
[G37] Nomada Studio . 2018. GRIS on Steam. Digital Game. https://store.steampowered.com/app/683320/GRIS/
[G38] Obsidian Entertainment . 2011. Fallout: New Vegas - Honest Hearts Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/fallout-new-vegas-honest-hearts/
[G39] Obsidian Entertainment . 2014. South Parkâ„¢: The Stick of Truthâ„¢ on Steam. Digital Game. https://store.steampowered.com/app/213670/South_Park_The_Stick_of_Truth/
[G40] One More Level, 3D Realms, Slipgate Ironworksâ„¢, All in! Games . 2020. Ghostrunner on Steam. Digital Game. https://store.steampowered.com/app/1139900/Ghostrunner/
[G41] Plastic Studios SCE Santa Monica . 2016. Bound (2016) Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/bound-2016/
[G42] PuffballsUnited . 2020. The Henry Stickmin Collection Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/the-henry-stickmin-collection/
[G43] Quantic Dream . 2020. Detroit: Become Human on Steam. Digital Game. https://store.steampowered.com/app/1222140/Detroit_Become_Human/
[G44] Quantic Dream . 2020. Heavy Rain on Steam. Digital Game. https://store.steampowered.com/app/960910/Heavy_Rain/
[G45] Radical Entertainment. 2012. Prototype 2 Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/prototype-2/
[G46] Recreate Games . 2023. Party Animals on Steam. Digital Game. https://store.steampowered.com/app/1260320/Party_Animals/
[G47] Red 5 Studios . 2014. Firefall Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/firefall/
[G48] Squanch Games, Inc. . 2019. Trover Saves the Universe on Steam. Digital Game. https://store.steampowered.com/app/1051200/Trover_Saves_the_Universe/
[G49] Squanch Games, Inc. 2022. High On Life on Steam. Digital Game. https://store.steampowered.com/app/1583230/High_On_Life/
[G50] Tango Gameworks . 2022. Ghostwire: Tokyo on Steam. Digital Game. https://store.steampowered.com/app/1475810/Ghostwire_Tokyo/
[G51] Telltale Games . 2012. The Walking Dead: A Telltale Games Series Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/the-walking-dead-a-telltale-games-series/
[G52] Telltale Games . 2012. The Walking Dead: Episode 5 - No Time Left Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/the-walking-dead-episode-5-no-time-left/
[G53] Telltale Games . 2016. The Walking Dead: The Telltale Series - A New Frontier Reviews - Metacritic. Digital Game. https://www.metacritic.com/game/the-walking-dead-the-telltale-series-a-new/
[G54] Tendershoot, Michael Lasch, ThatWhichIs Media . 2019. Hypnospace Outlaw on Steam. Digital Game. https://store.steampowered.com/app/844590/Hypnospace_Outlaw/
[G55] Tequila Works, QLOC . 2017. RiME on Steam. Digital Game. https://store.steampowered.com/app/493200/RiME/
[G56] The Astronauts . 2014. The Vanishing of Ethan Carter on Steam. Digital Game. https://store.steampowered.com/app/258520/The_Vanishing_of_Ethan_Carter/
[G57] the binary family . 2021. Save Farty â€“ the Trivia Game for Nintendo Switch - Nintendo Official Site. Digital Game. https://www.nintendo.com/us/store/products/save-farty-the-trivia-game-switch/?srsltid=AfmBOorcXS7Xw_x_sdJsVg1SELIHU_zGXrD9MVtNrsFWcU4m3qqR17Z9
[G58] Toys for Bob, Iron Galaxy Studios. 2019. Spyroâ„¢ Reignited Trilogy on Steam. Digital Game. https://store.steampowered.com/app/996580/Spyro_Reignited_Trilogy/
[G59] Tribute Games Inc. . 2022. Teenage Mutant Ninja Turtles: Shredder's Revenge on Steam. Digital Game. https://store.steampowered.com/app/1361510/Teenage_Mutant_Ninja_Turtles_Shredders_Revenge/
[G60] Tuttifrutti Interactive . 2017. Darkarta: A Broken Heart's Quest Standard Edition on Steam. Digital Game. https://store.steampowered.com/app/634180/Darkarta_A_Broken_Hearts_Quest_Standard_Edition/
[G61] Wishfully . 2023. Planet of Lana on Steam. Digital Game. https://store.steampowered.com/app/1608230/Planet_of_Lana/
[G62] GoodbyeWorld Games. 2021. Before Your Eyes on Steam. Digital Game. https://store.steampowered.com/app/1082430/Before_Your_Eyes/
[G63] Image & Form Games. 2022. The Gunk on Steam. Digital Game. https://store.steampowered.com/app/1087760/The_Gunk/
[G64] Storm in a Teacup. 2016. N.E.R.O.: Nothing Ever Remains Obscure on Steam. Digital Game. https://store.steampowered.com/app/377480/NERO_Nothing_Ever_Remains_Obscure/

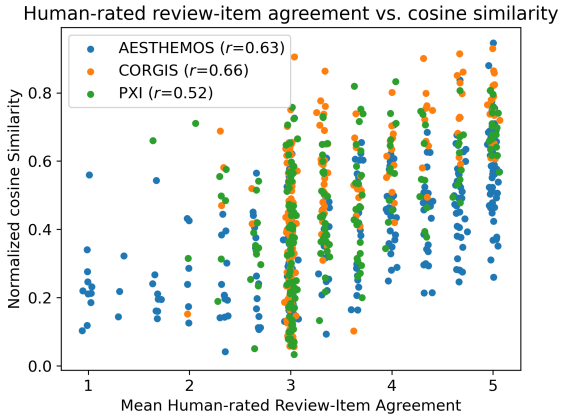


Figure 8: Scatterplot and Spearman correlations between review-item cosine similarities and mean human-rated review-item agreement.

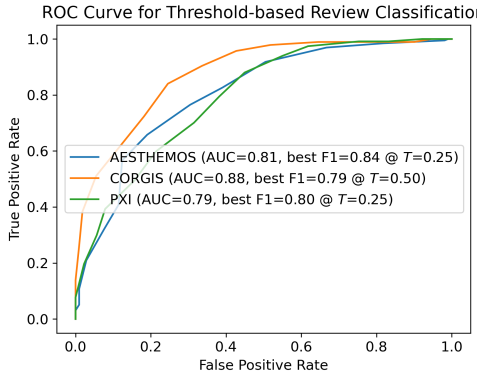


Figure 9: ROC curves, Area Under Curve (AUC), and best-case F1 scores and the corresponding thresholds (T) when using the thresholding to detect review-item agreement.

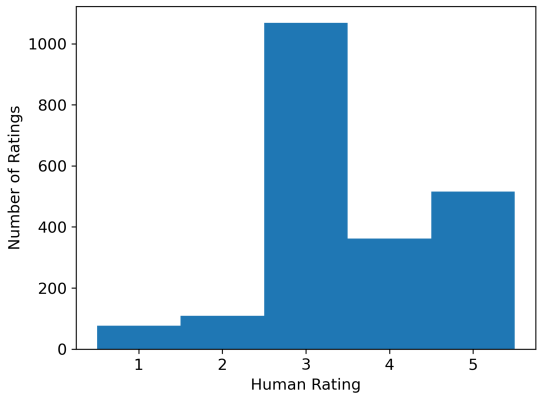


Figure 10: Histogram of all human-rated review-item agreements. The majority of ratings are neutral, confirming our assumption that in most cases, reviews only highlight one or a few aspects of player experience.

A Validation: Review-Item Cosine Similarity as an Approximation for Review-Item Agreement

To validate our use of embedding cosine similarity as a measure of review-item agreement, we compare the cosine similarities to human-rated review-item agreements, as detailed below.

A.1 Data

For each questionnaire item, we composed a stratified sample of 9 reviews ranging from the lowest to highest review-item cosine similarity. The stratification ensures that the sample is representative of both high and low similarities for each item. Given the 102 total items (42 for AESTHEMOS, 30 for, 30 for CORGIS) and the occasional case of a stratum containing no reviews, the combined sample size was 892 reviews.

Three authors manually coded the sample, rating the review-item agreements on a 5-point Likert scale (1: strongly disagree, 2: somewhat disagree, 3: neutral (not agree or disagree), 4: somewhat agree, 5: strongly agree). To minimize researcher bias, the order of the samples was randomly shuffled before coding and the embedding similarity values were hidden from the coders. Additionally, the coders flagged non-English reviews for removal, as English was the only shared language in which all the coders were proficient. The final coded sample of English reviews comprised 711 reviews.

To investigate intercoder agreement, we calculated the Spearman correlations between all 3 pairs of coders. The results indicate strong agreement (Pair 1: $r = 0.70, p < 0.0001$, Pair 2: $r = 0.77, p < 0.0001$, Pair 3: $r = 0.75, p < 0.0001$).

A.2 Notes on Data Quality

During coding, we noted that some reviews explicitly express agreement/disagreement with items like "Surprised me". However, one often needs to read between the lines. For instance, for the item "I felt indifferent", an example of a review that a human coder rated as disagreeing is “This was the best game experience ever.”

Regarding clear cases of sarcasm or otherwise untrue reviews, we spotted only this review for Disney’s Cory in the House: “I was in a difficult stage in my life, lost my job, lost my child, gained 400 pounds...I thought I was at point of no return but I found this game in a Bargain **** IT CHANGED MY PERSPECTIVE ON LIFE!! It motivated me to never give up and to follow my dreams, shed over 450 pounds and now i live in a big white house just like Cory.”.

We also noticed a total of 15 (2%) non-reviews such as “Hello worldHello worldHello worldHello worldHello...”, “fdc k s ksd k dkhs h jh...”, and “If you got a real mess and want to clean like the pros, then you’ve got to see this! Hi, Billy Mays here with Zorbeez, the most absorbent material...”. These were all

rated as neutral (3), but future work might employ an LLM to automatically remove non-reviews before further analyses.

A.3 Analyses

To validate that review-item cosine similarities are predictive of the human ratings, we calculated Spearman correlations between the cosine similarities and human ratings averaged over the coders. The results are shown in Figure 8 along with a scatterplot of the data.

To validate the threshold-based classification of reviews, we first created ground truth binary test data by labeling a review as "agree" if at least one human coder rated it 4 or 5, and "neutral/disagree" otherwise. This ground truth definition was adopted because during the manual coding, we found that determining review-item agreement was usually straightforward but the task was repetitive and lapses of attention may result in not spotting a part of a review indicating agreement, therefore increasing false neutral/disagree ratings. Using this ground truth data, we then analyzed the threshold-based classification performance using ROC curves and Area Under Curve (AUC), sweeping over thresholds in range [0,1] with a step of 0.05. The results are displayed in Figure 9. We further calculated the best-case threshold values for each questionnaire in terms of F1 scores. These are also reported in Figure 9.

To investigate the assumption that reviews only discuss one or a few salient aspects of player experience, suggesting that across all items, neutral review-item ratings should be the most common, we visualize a histogram of all the human ratings in Figure 10.

A.4 Results

As seen in Figure 8, the correlations for the cosine similarities are moderate to strong, depending on the questionnaire. This indicates that review-item cosine similarity is an approximation that is not perfect but should nevertheless produce meaningful results on a large dataset of reviews. The threshold-based classification analysis further confirms this. The AUC and F1 scores could be better but the AUC does indicate clearly better than random performance (which would yield AUC=0.5).

As expected, the histogram of all the human-rated review-item agreements in Figure 10 indicates that most of the agreement ratings are neutral, suggesting that a single review indeed only highlights one or a few key aspects of player experience. Furthermore, Figure 8 confirms that review-item cosine similarity has particularly high variance for the neutral reviews. This provides an explanation for trends not being visible in simple graphs of mean or median cosine similarities over time (Figure 3), motivating the thresholding as a means to focus on the strongest experiences.

A.5 Threshold selection

As noted above, it is possible to use the manually coded data to determine an "optimal" threshold for each questionnaire. In terms of the F1 scores above, such thresholds for AESTHEMOS, CORGIS, and PXI are 0.25, 0.5, and 0.25, respectively. However, *such a data-driven threshold selection is arbitrary in the end*, as it depends on the definition of the ground truth. For instance, if one defines the ground truth label for a review as "agree" if at least two coders rate it 5 (majority vote for strong agreement), the resulting thresholds are 0.45, 0.7, and 0.6.

For the trend visualization task of this paper, the high variance of the neutral similarity values suggests that one should rather use a too high than a too low threshold, lest the trends be drowned out by false positives. On the other hand, a too high threshold will result in so few reviews per year that the trend curves become noisy. The threshold values 0.45, 0.6, and 0.6 used in this paper lie within the two sets of "optimal" values above, and were finetuned manually to provide a compromise between low noise and high sensitivity.

A.6 Impact of review length

As longer reviews might be more confusing for the embedding model, resulting in lower similarity values, we calculated the Spearman correlation between review length and all review-item cosine similarities. The correlation is statistically significant but very weak ($r = 0.031$, $p<0.0001$). Therefore, the slight decline in review lengths over time is not likely to cause significant errors in our trend visualizations.

B Item-level Correlations for CORGIS Emotional Challenge

Table 6 shows the individual item-level correlations between emotional challenge item-review cosine similarities and review scores. The correlations exhibit considerable spread, with some items associated with high review scores, and others associated with low scores. This explains the overall weak correlation between emotional challenge and review scores.

C Experiment 3: Deductive Coding Instructions/Prompts

This section contains the LLM prompts used for for the deductive coding in Experiment 3. The LLM was prompted with the same prompt for each review, inserting it at the "<review>"

The human coders had all the reviews in a single spreadsheet and received the same instructions without the "Here is the review to code: <review>" part.

C.1 Emotional Challenge Prompt

Your task is to help in analyzing the reasons for experiencing emotional challenge in games, based on a dataset of game reviews.

Here, experiencing emotional challenge is conceptualized as an agreement with the following statements:

This game is more than just a game to me

The things that happened in the game made me sad

I invested much thought into the game

I felt a sense of responsibility for characters and events in the game

The game made me think about real life issues

Item	r	p
This game is more than just a game to me	0.30	0.000***
Playing the game was stimulating	0.29	0.000***
I felt a sense of suspense when playing the game	0.29	0.000***
I invested much thought into the game	0.23	0.000***
I felt a sense of responsibility for characters and events in the game	0.22	0.000***
The game made me think about real life issues	0.12	0.000***
The game had moral dilemmas in it where the choice was not obvious	-0.01	0.000***
The things that happened in the game made me sad	-0.32	0.000***
The game involved making moral choices that I didn't agree with	-0.39	0.000***

Table 6: Correlations (r) of review scores and review-item similarities for all CORGIS Emotional Challenge items. Interestingly, the items exhibit both positive and negative correlations. Due to our large sample size, even the weak correlations are statistically significant (p<0.001 denoted by *).**

Playing the game was stimulating

I felt a sense of suspense when playing the game

The game had moral dilemmas in it where the choice was not obvious

The game involved making moral choices that I didn't agree with

Precisely, we define emotional challenge as the average of the agreements with each individual statement. Thus, emotional challenge can manifest as a strong agreement with some statements, or a moderate agreement with many statements.

You should code the reviews deductively using one or more of the following codes, to indicate the reasons for the reviewer's experience of emotional challenge:

1. The game made me think about real life issues

2. Challenging narrative elements (themes, psychological, philosophical)

3. The game had moral dilemmas in it where the choice was not obvious

4. The game involved making moral choices that I didn't agree with (I could not choose differently)

5. I felt a sense of responsibility for characters and events in the game

6. The things that happened in the game made me sad

7. Other negative emotions, intentional (e.g., horror, suspense)

8. Other negative emotions, unintentional (e.g., bugs causing frustration)

9. Other reason

10. Not relevant (review does not express/discuss emotional challenge)

11. No reason (review does express/discuss emotional challenge, but reason for experiencing it is not clear)

Consider code "Not relevant" and other codes as mutually exclusive. If you code a review as "Not relevant", do not add other codes. If you add other codes, do not add "Not relevant".

Here is the review to code:

<review>

Please output the correct code numbers, without any text, separated by semicolons:

C.2 Boredom Prompt

Your task is to help in analyzing the reasons for experiencing boredom in games, based on a dataset of game reviews.

Here, experiencing boredom is conceptualized as an agreement with the following statements:

Bored me

Felt indifferent

Precisely, we define boredom as the average of the agreements with each individual statement. Thus, boredom can manifest as a strong agreement with one statement, or a moderate agreement with both statements.

You should code the reviews deductively using one or more of the following codes, to indicate the reasons for the reviewer's experience of boredom.

1. Repetitive/tedious/grindy

2. Unoriginal

3. Bad sequel, remake, or clone of some other game

4. Slow pacing

5. Lack of agency/choice

6. Lack of feedback

7. Bad/uninteresting story or dialogue

8. Bad/uninteresting level design

9. Bad/uninteresting audiovisual

10. Lack of depth

11. Lack of challenge/too easy

12. Good start turned boring

13. Low replayability

14. Tech/performance issues (e.g., loading times, bugs)

15. Other reason

16. Not relevant (review does not express/discuss boredom)

17. No reason (review does express/discuss boredom, but reason for experiencing it is not clear)

Consider code "Not relevant" and other codes as mutually exclusive. If you code a review as "Not relevant", do not add other codes. If you add other codes, do not add "Not relevant".

Here is the review to code:

<review>

Please output the correct code numbers, without any text, separated by semicolons:

C.3 Meaning Prompt

Your task is to help in analyzing the reasons for experiencing meaning in games, based on a dataset of game reviews.

Here, experiencing meaning is conceptualized as an agreement with the following statements:

Playing the game was meaningful to me

The game felt relevant to me

Playing this game was valuable to me

Precisely, we define meaning as the average of the agreements with each individual statement. Thus, meaning can manifest as a strong agreement with some statements, or a moderate agreement with many statements.

You should code the reviews deductively using one or more of the following codes, to indicate the reasons for the reviewer's experience of meaning.

1. Nostalgia
2. Story
3. Emotional impact
4. Motivational impact
5. Other personal impact (e.g., escapism during the covid pandemic)
6. Learned something from the game
7. Memorable
8. Thought-provoking
9. Overcame a difficult challenge
10. Unique game/pushes the boundaries/made me realize something about games
11. Historical/cultural immersion or references
12. Other reason
13. Not relevant (review does not express/discuss meaning)
14. No reason (review does express/discuss meaning, but reason for experiencing it is not clear)

Consider code "Not relevant" and other codes as mutually exclusive. If you code a review as "Not relevant", do not add other codes. If you add other codes, do not add "Not relevant".

Here is the review to code:

<review>

Please output the correct code numbers, without any text, separated by semicolons:

C.4 Nostalgia Prompt

Your task is to help in analyzing the reasons for experiencing nostalgia in games, based on a dataset of game reviews.

Here, experiencing nostalgia is conceptualized as an agreement with the following statements:

Made me feel sentimental

Made me feel nostalgic

Precisely, we define nostalgia as the average of the agreements with each individual statement. Thus, nostalgia can manifest as a strong agreement with one statement, or a moderate agreement with both statements.

You should code the reviews deductively using one or more of the following codes, to indicate the reasons for the reviewer's experience of nostalgia.

1. Childhood memories of the game or other games
2. First game/genre I played
3. Childhood memories of something else than a game (e.g., if the game is about fishing, remembering fishing with one's dad)
4. Nostalgic/retro graphics/sound/music
5. Sequel/prequel (evokes memories or comparisons of some original game)
6. Remake/recreation/mod (evokes memories or comparisons of some original game)
7. Rediscovering an old game/franchise, or coming back to gaming after a break
8. Big personal impact (career choice, preferences etc.)
9. Other reason
10. Not relevant (review does not express/discuss nostalgia)
11. No reason (review does express/discuss nostalgia, but reason for experiencing it is not clear)

Consider code "Not relevant" and other codes as mutually exclusive. If you code a review as "Not relevant", do not add other codes. If you add other codes, do not add "Not relevant".

Here is the review to code:

<review>

Please output the correct code numbers, without any text, separated by semicolons:

D Experiment 2: Contributions of Individual Genres and Games

This section contains the rest of the Experiment 2 stacked area charts, in Figures 11-20.

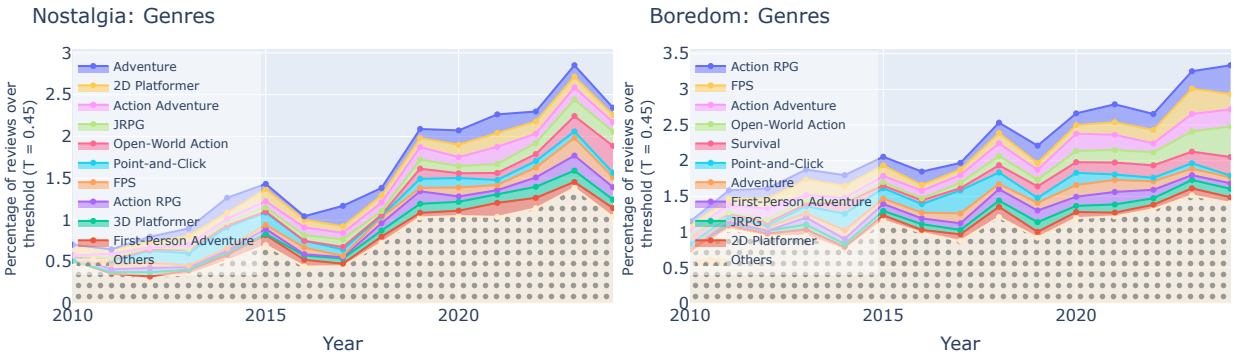


Figure 11: Top-10 genres for AESTHEMOS Nostalgia and Boredom

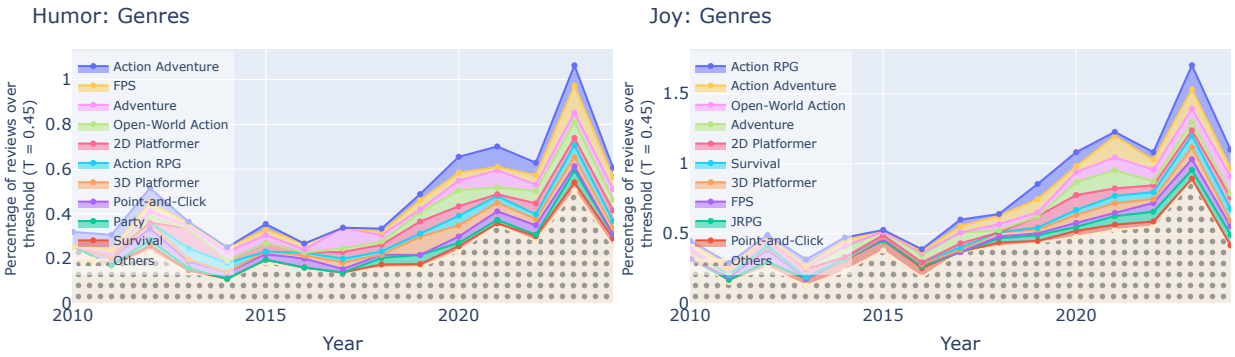


Figure 12: Top-10 genres for AESTHEMOS Humour and Joy

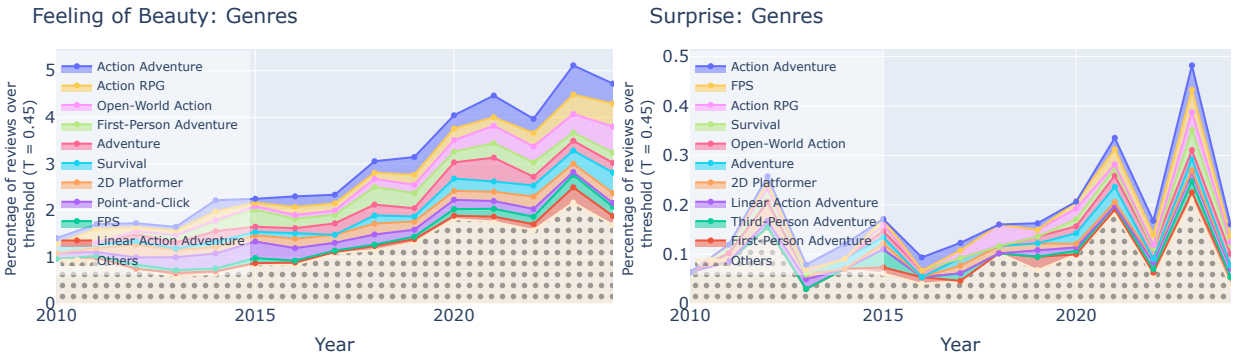


Figure 13: Top-10 genres for AESTHEMOS Feeling of Beauty/Liking and Surprise

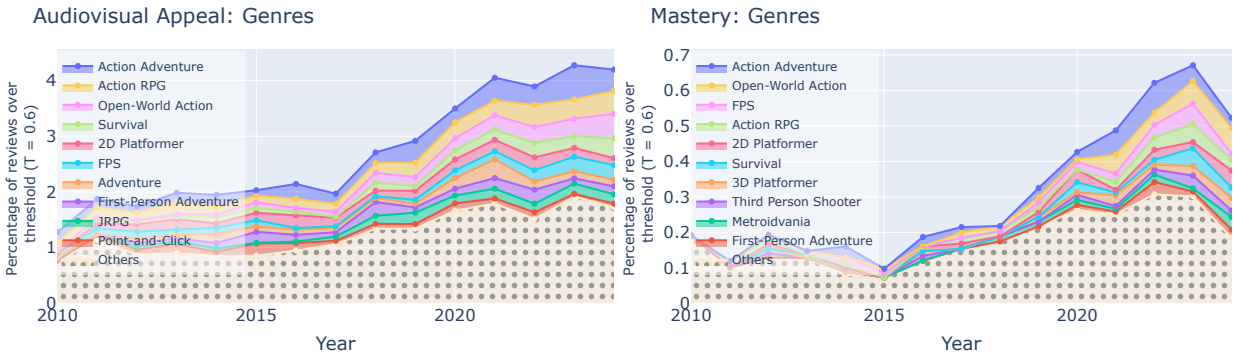


Figure 14: Top-10 genres for PXI Audio visual appeal and Mastery

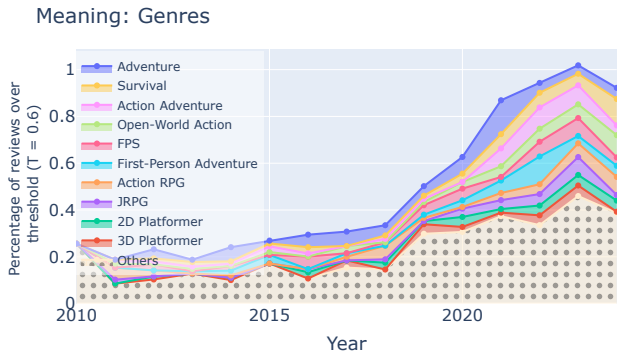


Figure 15: Top-10 genres for PXI Meaning

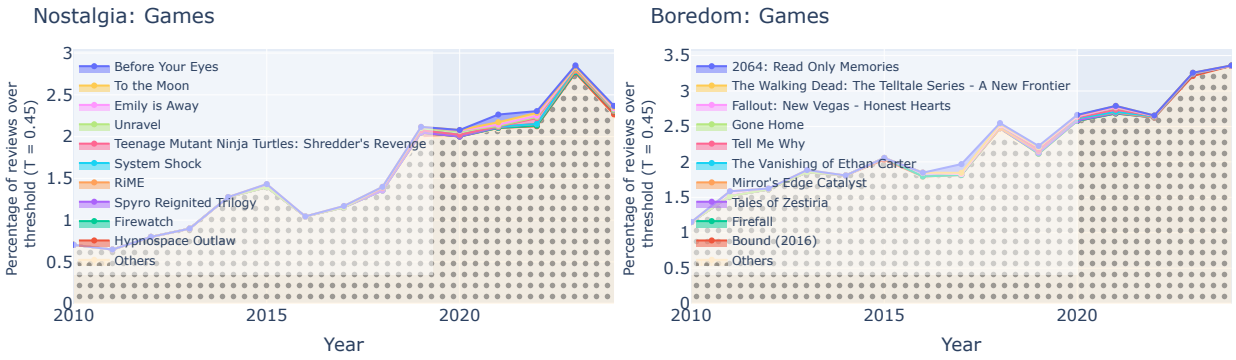


Figure 16: Top-10 games for AESTHEMOS Nostalgia and Boredom

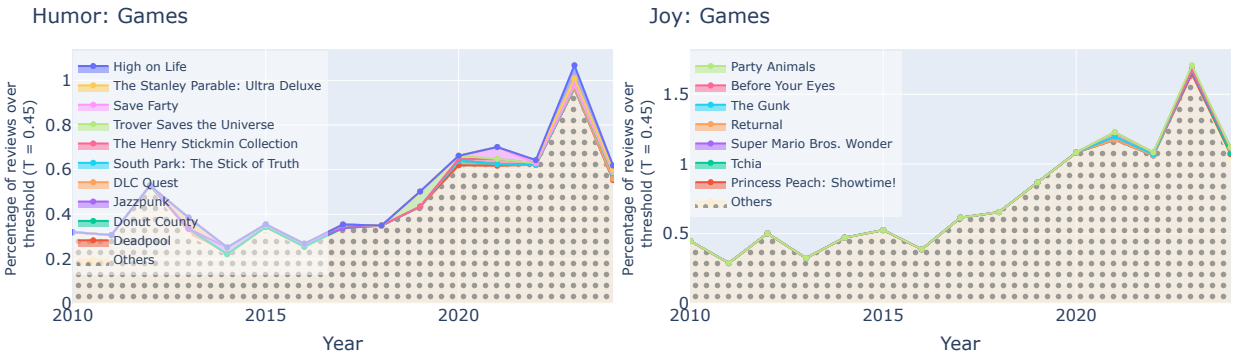


Figure 17: Top-10 games for AESTHEMOS Humour and Joy

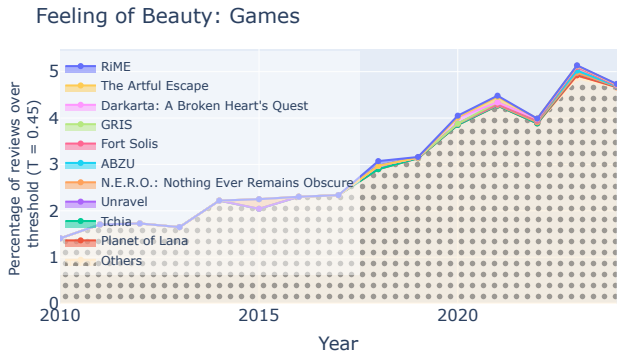


Figure 18: Top-10 games for AESTHEMOS Feeling of Beauty/Liking

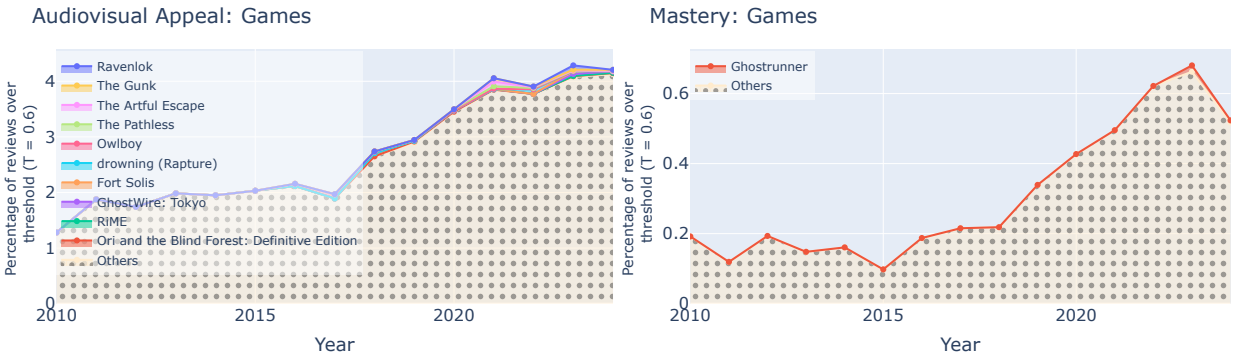


Figure 19: Top-10 games for PXI Audio visual appeal and Mastery

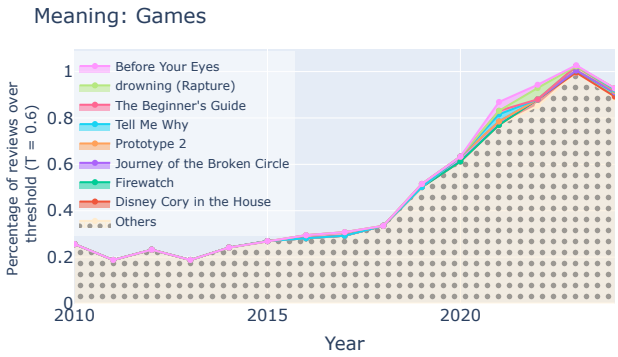


Figure 20: Top-10 games for PXI Meaning