İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025

https://doi.org/10.26650/acin.1670469
Submitted: 05.04.2025
Revision Requested: 11.04.2025
Last Revision Received: 14.05.2025
Accepted: 22.05.2025
Published Online 04.06.2025

Acta Infologica

Research Article

Open Access

Designing a Large Language Model-Based AI System for Dynamic Difficulty Adjustment in Digital Games

Onur Aşkın [1]

[1] Doğuş University, Faculty of Art and Design, Department of Digital Game Design, İstanbul, Türkiye

Abstract This study investigates how large language models (LLMs) can serve as dynamic agents in game-based interactions by comparing two prototypes of a color-guessing game. One model (Cohere Command) operates on a zero-shot prompt-based mechanism, while the other (FLAN-T5) is fine-tuned on a semantically structured dataset. A total of 20 participants were divided into two experimental groups to evaluate the models' ability to generate semantically coherent yes/no questions, maintain flow, and perform accurate predictions. Quantitative data, including session durations, number of interactions, and AI outputs, were analyzed, along with a post-game user experience survey grounded in Flow Theory. Results show that while both systems achieved task completion, the fine-tuned FLAN-T5 model significantly outperformed the other models in terms of semantic clarity, user engagement, and perceived fluency. The findings highlight the potential of LLM-based DDA systems in creating meaningful, adaptive player experiences and underscore the importance of semantic alignment and interaction transparency in game-based AI design.

Keywords Artificial Intelligence · Machine Learning · Digital Game Design · Dynamic Difficulty Adjustment

Citation: Aşkın, O. (2025). Designing a large language model-based AI system for dynamic difficulty adjustment in digital games. *Acta Infologica*, Advance Online Publication. https://doi.org/10.26650/acin.1670469

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Aşkın, O.

Corresponding author: Onur Aşkın oaskin@dogus.edu.tr

Introduction

When a game's difficulty level does not match the player's skill level, the experience becomes tedious or overly challenging. To achieve this balance, Dynamic Difficulty Adjustment (DDA) systems aim to make the game flow more motivating and sustainable. Traditional DDA approaches are usually based on numerical data such as the player's success rate and score. However, these methods do not adequately consider a player's decision-making style, contextual strategies, and in-game interaction patterns.

Integrating big language models (LLMs) into game systems provides a new dimension to the concept of LDA. Such models enable the development of AI units that are not only data-driven but also able to read context, analyze previous interactions, and make strategic decisions (Brown et al., 2020; Radford et al., 2019). Thanks to the flexible nature of the language, AI can offer player-specific difficulty levels and enable more human-like decisions. Game scenarios with short yes/no answers provide a suitable ground to test the contextual analysis power of language models. This kind of experimental setup allows us to observe how AI performs at different difficulty levels while rethinking the role of these systems in game design.

However, in the existing literature, the use of LLMs in the context of dynamic difficulty tuning has been addressed in a limited number of studies; in particular, the effects of different LLM architectures on in-game strategic decision making, semantic output quality, and flow experience have not been systematically compared. Aiming to fill this gap, this paper empirically compares two different LLM architectures (zero-shot and fine-tuned) in a game-based interaction setting and presents a multi-layered evaluation of semantic directiveness, flexibility, and user experience metrics.

Purpose of the Study

This study investigates how artificial intelligence using large language models (LLMs) contributes to adjusting the difficulty level of games according to the player. For this purpose, a question-answer-based guessing game between two artificial intelligences was designed, and their performances were compared at different difficulty levels: easy, medium, and hard.

Research Question and Hypotheses

Research Question: Can large language models optimize dynamic difficulty adjustment in digital games with humanlike strategic decision-making?

Hypothesis 1 (H1): AI-assisted dynamic difficulty adjustment (DDA) systems can be optimized using large language models (LLM).

Sub-Hypotheses:

- H1a: LLM-based prediction strategies enable AI systems to dynamically adjust in-game difficulty levels in real time.
- H1b: LLM-supported AI systems can make strategic decisions at different difficulty levels (easy, medium, hard).
- H1c: LLM-based AI predictions show lower response times and higher decision flexibility than traditional machine learning algorithms.
- H1d: The output entropy of LLMs increases the accuracy and adaptability of AI-based prediction strategies.

- H1e: LLM-based DDA systems provide more consistency in game flow and player interaction than traditional methods.

Hypothesis 0 (H0): LLM-based AI systems do not provide significant improvements in terms of dynamic difficulty adjustment.

Variables

Independent Variables:

- Natural Language Processing (NLP) Model: Language processing capacity that drives AI's prediction strategies.
- Difficulty Level (Easy, Medium, Hard): The complexity level of the AI prediction strategies.
- Prediction Time: The prediction time taken by the AI algorithm.
- Success Rate: The rate at which the AI correctly assumes the hidden color.

Dependent Variable:

- Prediction Strategy Success: The proportion of correct predictions based on the language model used by AI in dynamic difficulty adjustment strategies.

Scope and Limitations

This research used an AI-assisted color prediction game to evaluate the performance of large language models (LLMs) in dynamic difficulty adjustment (DDA) systems. After playing the game, the participants completed a user experience survey with questions generated by two different artificial intelligence models (Cohere AI and FLAN-T5) at different difficulty levels. Thus, a multi-layered evaluation was performed in line with the system data and user opinions.

However, the study has certain limitations.

Within the scope of the research, only two LLM architectures were examined; models from other platforms, such as OpenAI, IBM, and Azure, were not included in the study. The Cohere model was used as a routing-based model from abrasion, while the FLAN-T5 model was fine-tuned and implemented on the Hugging Face platform with a custom dataset. However, different model versions, advanced training configurations, or alternative task protocols offered by Hugging Face were excluded from the scope of this study. In the implemented models, the AIs were not allowed to learn the difficulty levels independently; the difficulty settings were determined according to predefined rules.

Data analysis was limited to key performance indicators such as accuracy rate, number of questions, and playing time; however, no advanced methods such as long-term learning processes, time series analysis, multiple user interactions, or in-game social dynamics, were used. Interactions were text-based, with no visual or auditory supportive components.

Ethically, no personal data were collected in the study; all sessions were conducted based on participant consent and volunteerism. The research was conducted with the approval of Doğuş University Scientific Research and Publication Ethics Committee (Date: 28.05.2025, No: 82320). When similar systems are planned to be tested with real player data in the future, basic ethical principles, such as data privacy, user consent, algorithmic transparency, and system auditability, must be rigorously observed.

Methodological Justification and Design Rationale

Although the number of participants in this study (n = 20) seemed limited, the experimental design was deliberately divided into two independent groups (n_1 = 10, n_2 = 10). This partitioning eliminates factors that threaten the internal validity of comparing AI systems, such as the learning effect, ranking effect, and surrogate learning. The experimental design ensured that all participants interacted with only one model, preventing cross-contamination and allowing for a more precise effect size measurement.

Furthermore, another prominent aspect of this study is the multi-layered data collection and analysis strategy. Different layers of data were collected and analyzed, including game session data, user experience surveys, Shannon Entropy analyses, and independent sample t-tests for statistical significance. This increased the study's contextual depth and validity despite the quantitative sample's limitations.

Another study strength is that the game prototypes are reproducible through open-source platforms. This ensures the experiment's replicability and provides a scalable frame of reference for future work. Therefore, the study goes beyond a pretest and presents a transparent model in terms of experimental design and offers added value for practice.

This methodological justification section provides defensible arguments that emphasize the empirical nature of the study, data depth, and analyzability despite superficial criticisms about the number of participants. On the other hand, only two major language models (Cohere Command and FLAN-T5) were compared in this study; other models, such as OpenAI GPT, Claude, and Gemini, were excluded. This choice was made deliberately to ensure methodological control of the study and to perform an in-depth comparison. The selected models represent different production approaches: Cohere Command offers a prompt-based architecture, while the FLAN-T5 model has a fine-tuned architecture with a specialized dataset. Thanks to this structure, we can directly compare the advantages and limitations of different language model-based dynamic difficulty tuning methods. In similar studies, the number of models is usually limited, and factors such as methodological consistency, parametric control, and technical resource requirements are considered in comparative analyses. In future research, comparisons involving more language models can be performed to increase generalizability and analyze different architectures' production behavior more comprehensively.

Theoretical Framework

This study's theoretical basis is the Flow Theory developed by Csikszentmihalyi (1990). Flow is a special state of consciousness in which the individual fully focuses their attention on the activity loses the perception of time, and experiences high motivation during the activity (Keller & Bless, 2008). In order to sustain this experience in game design, an ideal balance between the player's skill level and the game's difficulty is recommended.

Sweetser and Wyeth (2005) highlighted three key elements that support the flow state: clear goals, immediate feedback, and skill-appropriate challenges (Cowley et al., 2008). These elements strengthen the player's sense of control and emotional connection to the game, thus satisfying the experience. Chen (2007) took these principles to an algorithmic level and developed the concept of "Dynamic Difficulty Adjustment" (DDA). DDA systems aim to maintain the continuity of the flow state by changing the game difficulty in real time according to the player's performance.

Difficulty setting and balancing cognitive load are critical for sustaining the Flow experience. This is where Sweller's (1988) Cognitive Load Theory comes into play. Systems that allow the player to make meaningful decisions within the game without interrupting their attention with excessive stimuli increase the quality of learning and experience.

Ryan and Deci's (2000) proposed the self-determination theory, which provides an important structure that supports the flow in a motivational framework. According to this theory, intrinsic motivation increases when individuals' basic psychological needs, such as competence, autonomy, and relationship building, are satisfied. Dynamic challenge systems sustain this motivation to the extent that they provide the player with a sense of competence.

Unlike classical rule-based LLM frameworks, artificial intelligence systems based on big language models (LLM) not only generate responses to player inputs and have the capacity to develop semantic contextual guidance, meaningful question generation, and prediction strategies (To & Brusilovsky, 2021).

In this study, we compare two different big language models, prompt-based (Cohere AI) and fine-tuned (FLAN-T5), and evaluate their impact on semantic production quality, guidance capacity, and flow experience. The findings revealed significant differences in technical accuracy and experiential parameters, such as player perception, time perception, and cognitive continuity. In this respect, this study proposes a multi-layered rethinking of how dynamic difficulty adjustment systems can be made semantically guided and personalized at the individual interaction level through LLM-based production behaviors in the context of flow theory.

Literature

This section provides a literature review on Dynamic Difficulty Adjustment (DDA) systems, structured around several key themes: their application in games, their impact on player experience, integration with artificial intelligence (AI) and machine learning (ML), personalization strategies, and their use in education and rehabilitation. In addition, this review highlights a significant gap in studies that integrate Large Language Models (LLMs) with DDA systems. This review aims to establish a foundational understanding of DDA's technical and experiential dimensions by outlining these domains.

DDA refers to techniques that dynamically adjust game difficulty in response to the player's performance, skill level, or emotional state. This adaptive approach is often employed to support player motivation and facilitate flow experiences. Early studies, such as Gilleade et al. (2004), examined how adaptive difficulty could sustain players in a flow state, while Kowlessar (2020) demonstrated that optimal difficulty settings enhance player engagement. However, some findings suggest that relying solely on performance metrics is insufficient; psychological states must also be considered to avoid unintended player frustration. The application of AI and ML techniques has significantly advanced the design of DDA systems. For instance, Garcia-Ruiz et al. (2023) introduced a deep learning-based system that increased player satisfaction by adjusting difficulty based on skill progression. Roohi et al. (2021) combined Deep Reinforcement Learning (DRL) with Monte Carlo Tree Search (MCTS) to create an adaptive system that performs well at higher difficulty levels. Similarly, Yannakakis and Togelius (2018) and Lopes and Bidarra (2011) emphasized the central role of adaptive AI in dynamic gameplay experiences. Fisher and Kulshreshth (2024) contributed to this discussion by demonstrating that performance- and emotion-based DDA strategies produce varying outcomes across player profiles. In addition to performance optimization, DDA systems have been employed in motivational and cognitive frameworks. For example, Colwell et al. (2018) demonstrated that adaptive

difficulty calibrated using the ARCS motivational model improved trust and engagement. Smulter and Porsbjer (2019) examined how cognitive workload affects players under dynamically changing difficulty, suggesting that task complexity and attentional demand should be central parameters in adaptive systems. The growing interest in personalization has led to emerging studies using large language models (LLMs) in DDA. However, work on integrating LLMs with DDA remains limited, revealing a crucial research gap. In applied contexts, DDA has also shown potential in educational and therapeutic settings. Wu, Chen, and Chen (2017) demonstrated that an adaptive e-learning system based on dynamic scaffolding significantly enhanced student learning performance by aligning difficulty levels with individual progression. Cameirão et al. (2009) highlighted the use of the Rehabilitation Gaming System (RGS) as an example of how adaptive game-based environments can support cognitive and motor recovery in clinical neurorehabilitation. From a technical standpoint, Zheng (2024) reported that the integration of deep reinforcement learning into DDA systems enhanced the adaptability and precision of difficulty scaling, enabling more strategic and skill-responsive level design. These insights reinforce the role of DDA in structuring challenges in a way that aligns with individual player trajectories. In conclusion, the literature collectively demonstrates the effectiveness of DDA systems in terms of enhancing gameplay, motivation, and learning. However, integrating LLM-based AI systems into DDA remains an underexplored area. Although personalization through AI-driven adaptation has been considered in recent work, existing research still lacks comprehensive, comparative, and experimentally validated frameworks that assess how different LLM architectures influence dynamic difficulty adjustment across gameplay variables. In particular, little is known about the strategic decision-making capacity, semantic output quality, and flow facilitation of LLMs under different difficulty tiers. This study addresses this critical gap by experimentally comparing a zero-shot prompting-based model (Cohere Command) with a fine-tuned LLM (FLAN-T5) in a controlled, game-based interaction context. By examining semantic richness, adaptability, and user engagement metrics, the research aims to provide a multilayered evaluation of how LLM-driven DDA systems can be optimized for more intelligent and humanlike gameplay experiences.

Methodology

Experimental Structure and Setup

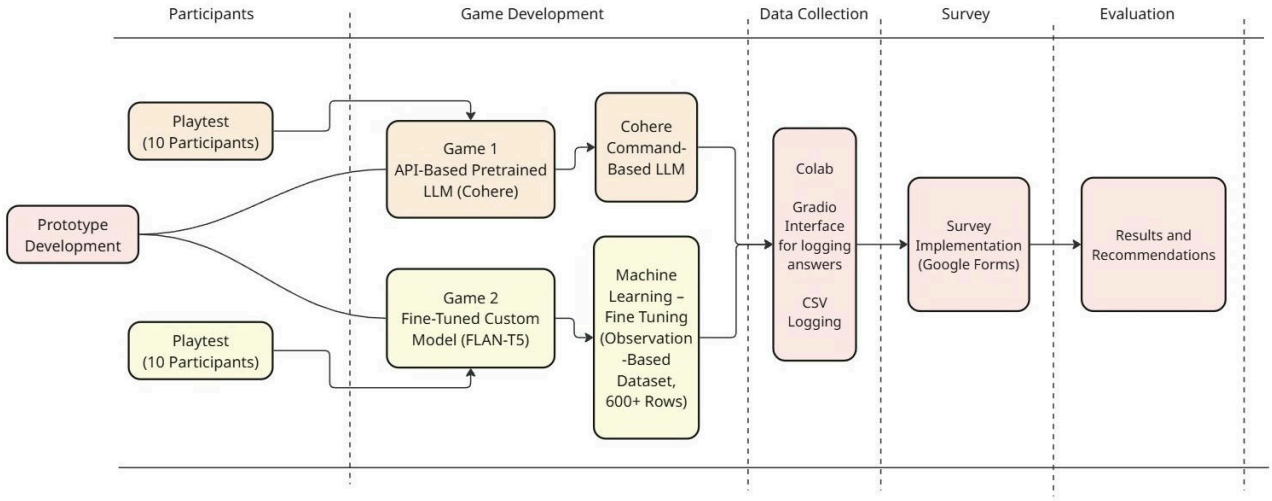
This study used an experimental method to examine the behavior, semantic, and experiential performance of large language models (LLMs) in game-based interactions. Within the scope of the research, two different game prototypes were developed, and the plar in their minds through yes/no questions. The first of these prototypes is based on the zero-shot prompting technique with the Cohere AI model, while the second one is fine-tuned and customized on the FLAN-T5 architecture on the Hugging Face platform. Both games were structured with three different difficulty levels (easy, medium, and hard), and the participants' responses to the semantic prompts were evaluated. Participants were divided into two groups. Each group experienced only the game prototype directed by an artificial intelligence model. In this way, internal validity issues that may be observed in comparative experiments, such as learning and ranking effects, were avoided (n_1 = 10, Cohere; n_2 = 10, FLAN-T5).

Accessibility, architectural diversity, and the feasibility of fine-tuning were considered in the model selection. The FLAN-T5 model stands out due to its open-source nature, easy trainability on the Hugging Face platform, and semantic generation success in the literature. The Cohere Command model, on the other hand, is known for its zero-shot performance based on routing and provides a suitable counterexample for comparative testing in terms of natural language generation without the need for fine-tuning. The fact

that both models represent different production approaches provides a meaningful framework for the experimental comparison of their production behavior and the analysis of their impact on dynamic difficulty management (Figure 1).

Figure 1

Experimental Workflow of AI-Supported Game Prototype Development and Evaluation



Technical Implementation and Game Prototypes

The Cohere AI-based game version is based on the Command model, accessed via a ready-made API, and driven only by system prompts. The model was expected to generate short questions focusing on color properties that could only be answered with yes or no answers. At the three difficulty levels, production variance and semantic complexity gradually increased. The FLAN-T5 model was fine-tuned on a custom dataset comprising more than 600 color-based question-answer pairs. This process was run on Google Colab Pro in the Hugging Face environment using the Transformers and Seq2SeqTrainer classes (Vaswani et al., 2017). The model was optimized for strategic question generation by considering factors such as semantic context, psychological connotation, and color categories (Goodfellow, Bengio, & Courville, 2016).

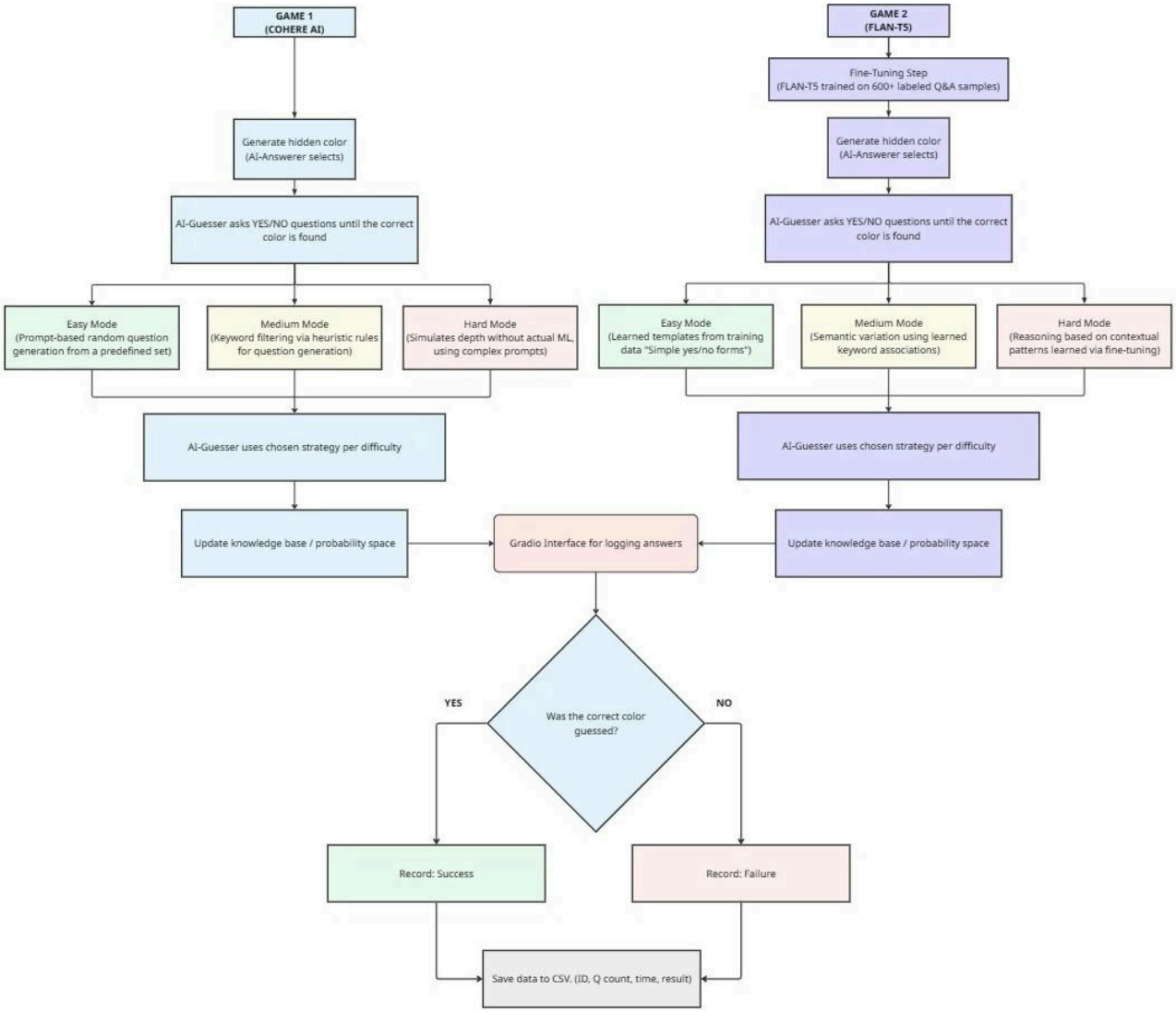
Both games were presented to the user through a Gradio-based interface, and all sessions were recorded throughout the production process. In the color prediction process, the system had the user predict a fixed target color; thus, the artificial intelligence was directed to make meaningful inferences based only on user responses.

The training dataset used in the fine-tuning process of the FLAN-T5 model was created using a large language model (ChatGPT) that supported text generation according to the semantic rules created by the researcher. The dataset contains over 600 yes/no questions in total, each labeled with a target color, difficulty level (easy, medium, hard), and semantic category (e.g., nature, temperature, symbolism). For example, the question "Is this color often linked with temperature in nature?" is in the category "temperature" under the color "burgundy" and the difficulty level "Easy". In the production process, multi-layered conceptual areas, such as the hot-cold perception of colors, the frequency of their occurrence in nature, cultural connotations, and emotional effects, were considered. The researcher manually reviewed the generated data samples, and the reliability of the training set was ensured by removing content containing semantic

inconsistencies or physical meaning errors. The full dataset is presented in the appendix of this article (Figure 2).

Figure 2

Comparative Flowchart of Two LLM-Based Color Guessing Games



Data Collection, Survey, and Analysis Process

Interaction data were saved in CSV format during all game sessions. These data includes session IDs, question texts asked by the AI, user responses, remaining color probabilities, and prediction times.

At the end of the game, each participant was administered a Likert-type user experience questionnaire designed in accordance with Flow Theory. The questionnaire aimed to measure experiential elements such as AI question clarity, content repetition rate, prediction success, time perception, and overall immersion.

The collected data were analyzed using Python. The following libraries were used in this analysis: pandas, seaborn, scipy, and collections (Harris et al., 2020; Virtanen et al., 2020). The key metrics computed include the semantic repetition rate, number of out-of-context productions, session duration, and entropy. The performance comparison between models was performed using independent samples t-tests.

Entropy Analysis: Diversity in User Opinions

The user survey data were evaluated not only in terms of average scores but also in terms of the diversity of responses. For this purpose, Shannon Entropy calculations were performed. The entropy is particularly meaningful for assessing the subjective dimensions of the player experience. The more diverse user responses to a metric, the more flexible and multilayered the interpretation of that experience metric.

The entropy value is calculated using the following formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot log_2 p(x_i)$$

Here$p(x_i)$denotes the probability of each response category (Likert options 1-5) being observed. The higher the entropy value, the more diverse the respondent responses are; the lower the entropy value, the more homogeneous are their responses.

These calculations were performed using collections. Counter and math.log2 functions in Python (Harris et al., 2020). Thus, user experience was evaluated not only in terms of score averages but also in terms of differences, level of unpredictability, and variety of cognitive awareness.

Evaluation of Outlier Duration Values

When the interaction data of the Game 1 (Cohere AI) game prototype were analyzed, it was found that there were some unusual (outlier) values, especially in session durations. For example, some sessions exceeded 135,000 seconds (approximately 37 hours), which is not a realistic interaction duration in the context of the experimental task. Upon further investigation, it became clear that such data were due to the game being left open (idle state) without the users actively participating.

This is a common problem, particularly in web-based experimental studies. The participant can walk away from the computer, tab to the background, or do something else without closing the game screen. The game interface does not close during such passive situations; thus, the system continues to record the session duration without interruption. In this context, in this study conducted on the Gradio interface, long-opening the browser tab was recorded as game time; however, these values do not reflect the actual user interaction.

To avoid this problem and ensure the validity of the data, interaction time alone was not used as an evaluation criterion; instead, we focused on more reliable indicators, such as the number of questions, production entropy, and user experience surveys. Furthermore, only active and consistent interaction patterns were included in the statistical analysis, ignoring such outliers. This approach ensured that the results accurately represented the interaction between the user and the AI.

Results

In the process of evaluating the game prototypes developed with two different large language models, both in-game interaction data and survey results based on user tests were analyzed in multiple layers. The playability performance of both games was measured using CSV-formatted game data obtained during plates sessions conducted with real users; AI-generated questions, player responses, and processing times were recorded in detail. These in-game data were systematically analyzed in terms of the AI model's guidance, production consistency, and prediction strategy. Furthermore, cognitive, semantic, and flow feedback on the game experience was collected through structured questionnaires that the participants

answered after each game session. By evaluating the quantitative and qualitative data sources together, the findings presented below were reached, and the performance characteristics of both models were presented comparatively in the context of game interaction.

Game 1 (Cohere AI)

The interaction data of 10 individuals participating in the color guessing game were analyzed at three different difficulty levels (easy, medium, hard) based on the total game time and the number of questions generated. This analysis provides important findings on the productivity, semantic guidance capacity, and response generation tempo of AI-supported interactions.

Table 1

Detailed Interaction Time and Question Count Across Difficulty Levels in Game 1 (Cohere AI)

Participant ID	EASY		MEDIUM		HARD		TOTAL	
	SEC	Q COUNT	SEC	Q COUNT	SEC	Q COUNT	SEC	Q COUNT
1	1636.79	30	13802.49	158	420.72	22	15860.00	210
2	297.81	39	135772.98	112	1237.32	63	139982.11	214
3	628.17	21	32390.66	189	12321.56	79	45340.39	289
4	1177.12	20	19103.48	161	1243.99	34	21524.59	215
5	2590.70	39	10868.40	124	4336.61	61	17795.71	224
6	157.43	12	106001.85	220	2564.08	43	108723.36	275
7	4623.27	39	159620.68	310	2826.37	50	167070.32	399
8	1329.26	24	39339.04	212	4672.47	46	45340.77	282
9	1867.80	30	214292.23	199	21295.94	91	237455.97	320
10	815.26	21	147767.02	316	3675.33	28	152257.61	365

When the overall distribution is analyzed, it can be seen that the sessions at the medium difficulty level lasted significantly longer (M = 1793 minutes, SD = 990). In contrast, the average duration at the easy level was 38.7 minutes. The difficult level was positioned between the two levels with an average of 106.9 minutes. These differences suggest that interactions with large language models are strongly influenced by technical parameters and contextual stability (Table 1).

Table 2

Summary of Average Duration and Question Count by Difficulty Level (in Minutes)

Difficulty Level	Average Duration (min)	Standard Deviation (min)	Average Question Count	Standard Deviation (count)
Easy	38.69	24.18	29.6	9.2
Medium	1793.28	990.34	202.0	67.0
Hard	106.95	97.17	51.7	22.1

There was a similar disproportion in terms of the number of questions. An average of 29.6 questions were produced in the easy mode; this number reached 202 in the medium mode. This dramatic difference suggests that the system has entered a kind of "contextual stuck" state. At the intermediate level, the model continued to generate questions without making any meaningful predictions, which caused the guiding interactions to be replaced by aimless production. The basis of this situation is that the model cannot end the game at any point and cannot make a prediction, such as "Is this color red?" to the user. The fact that

the standard deviation at this level reached 990 minutes supports the observation that in some sessions, the model exhibited serious inconsistencies in both semantic and temporal terms. As a matter of fact, in some examples, the model's generating questions contrary to the rules of the game such as "Is this color integrated with nature?" or "Does this color change according to the seasons?" caused the players to become indecisive and prolonged the interaction time (Table 2).

Furthermore, hallucinatory (unreal, contextually irrelevant, or physically contradictory) content generation was observed in all participants throughout the interaction process. Although grammatically coherent, the content did not provide meaningful guidance in terms of the game context; on the contrary, it interrupted the flow of the game experience by increasing the users' cognitive load. Similarly, physically contradictory questions such as "Is this color both warm and blue?" revealed that the model could not perform conceptual fact-checking and only generated content based on statistical patterns. The prevalence of such production indicates an overproduction and a serious lack of semantic control in the quality of production.

In hard mode, the number of questions was relatively limited (M = 51.7 questions, SD = 22.1), and the generation time was shorter and more consistent (M = 106.9 minutes, SD = 97.2 minutes) (Table 2). This may indicate that the AI is trying to execute production more carefully when faced with more complex prompts. However, this positive trend does not eliminate the system's tendency to fall again. For example, asking three questions with the same meaning in succession, such as "Is this color warm?" "Is it one of the warm tones?" and "Is it close to red?" clearly demonstrate the model's inability to use short-term context memory effectively.

In order to stabilize the production process of the model, a short-term micro-context memory was created, and the previous three questions were included in the prompt. However, this strategy was not statistically effective because of the context window limitation. It was systematically recorded that similar questions were repeated many times within the same session, meaningless jumps were made, and questions irrelevant to the task were produced. This, coupled with the non-deterministic production behavior of the model, led to a pattern in which uncontrolled variance was transformed into semantic distortion. In particular, the generation of different and often non-directional responses to the same prompt, despite the temperature parameter being fixed at 0.5, demonstrates the inherent inconsistency of the production. The hallucinatory productions not only prevented the natural ending of the game but also caused the players were unable to understand what the AI was aiming for.

These findings demonstrate that the production behaviors of LLM-based AI systems operate solely based on linguistic patterns and have significant limitations in contextual task adaptation. When the model operates without a specific contextual goal, it tends to engage in uncontrolled production, which directly impacts user experience. In such cases, even if the amount of production is high, the content remains weak in terms of guidance and semantic accuracy, and the semantic coherence of the system is sacrificed in favor of productivity. Production "abundance" was often inversely proportional to production "quality".

Large language models appear vulnerable to exhibit directionless and uncontrolled behavior in free production structures and game-like interaction environments. This behavior threatens not only the user experience and the logical flow of the game. Therefore, in the in-game use of LLM systems, the outputs and production process must be controlled and redesigned in a multi-layered manner. Otherwise, the players' experience will be semantically unsatisfactory, technically inefficient, and repetitive.

Game 1 Survey Evaluation

This study collected user experience data through a 5-point Likert-type questionnaire administered to participants who interacted with an AI-powered color guessing game (n = 10) (Table 3).

Table 3

User Experience Survey Results for Game 1 (Cohere AI)

Question	Number of Participants *(n)*	Mean *(M)*	Standard Deviation *(SD)*	Minimum *(Min)*	1st Quartile *(Q1)*	Median *(Md)*	3rd Quartile *(Q3)*	Maximum *(Max)*
I immediately understood the questions asked by the AI.	10	2.1	1.449	1.0	1.00	1.5	2.75	5.0
Similar questions were repeated repeatedly.	10	4.4	1.264	1.0	4.25	5.0	5.00	5.0
The AI tried guessing the color.	10	1.5	0.527	1.0	1.00	1.5	2.00	2.0
I lost track of time while playing the game.	10	1.4	0.699	1.0	1.00	1.0	1.75	3.0
The game was very engaging.	10	1.4	0.699	1.0	1.00	1.0	1.75	3.0

Participants' responses were interpreted in terms of the system's comprehensibility, content diversity, predictive capacity, and experiential immersion, and the statistical data obtained were supported by qualitative content and subjected to analysis.

"The mean M = 2.10, standard deviation SD = 1.45, median Md = 1.5, minimum = 1, maximum = 5, first quartile ($Q1$) = 1.00 and third quartile ($Q3$) = 2.75. This result indicates that a large proportion of the participants had difficulty understanding the AI questions. However, the high standard deviation reveals that there is significant variation in this perception among the participants. In this context, the fact that some participants had difficulty understanding the questions directly can be attributed not only to the semantic inadequacy of the AI but also to the naturalness of the form of questioning offered by the system, i.e., its capacity to engage in human-like interaction. Rather than a technical failure, such a "problem of comprehension" may be due to the artificial system producing a sufficiently humanoid form of language that renders its artificiality invisible. This reflects the success of large language models (LLMs) in interactive systems (Table 3).

The findings regarding the statement "Similar questions came over and over again." are M = 4.40, SD = 1.26, Md = 5.0, Min = 1, Max = 5, $Q1$ = 4.25, *and* $Q3$ = 5.00. The fact that 75% of the participants responded positively to this item creates the impression that the questions posed by the system are not sufficiently diversified in terms of content and fall into repetition at the perceptual level. The fact that 75% of the participants chose the two most positive options (4 and 5) indicates that the questions posed by the artificial intelligence have paternal similarities, and users can easily recognize this situation. This finding indicates that the system must improve semantic depth and variation generation. However, it should be noted that repetitive questions should not always be considered negative, as this may also be due to the fact that the system intentionally posed similar questions in a controlled manner while generating predictive hypotheses (Table 3).

The mean of the participants' responses to the statement "Artificial intelligence tried to guess the color" was found to be relatively low (*M* = 1.50, *SD* = 0.53, *Md* = 1.5, *Min* = 1, *Max* = 2, *Q1* = 1.00, *Q3* = 2.00). This value indicates that most participants either did not notice the AI's prediction behavior or the system failed to present this function clearly and distinctly enough. Thus, there is a serious lack of transparency in terms of how the prediction process is reflected to the user. The player's ability to intuit how the system works, from which information it makes predictions, and when it switches to generating a response is critical to the credibility of the artificial agent with which they interact. Therefore, the prediction process should not only be algorithmic but also be supported by clear and visible experiential feedback to the user.

Responses to the statements "I did not realize how time passed while playing the game." and "The game was very immersive." were evaluated with equal scores as *M* = 1.40, *SD* = 0.70, *Md* = 1.0, *Min* = 1, *Max* = 3, *Q1* = 1.00, *and Q3* = 1.75, respectively. These low averages indicate that users could not enter a flow state while experiencing the game and were not sufficiently involved in the game in terms of attention and time perception (Kaye, 2016). When evaluated within the framework of flow theory, the lack of cognitive intensity revealed that the system was limited in immersing the user and creating a sense of continuity. In particular, the low immersion and time perception parameters indicate that the game's interactional dynamics or esthetic experience is not at a level that satisfies the player.

It is also noteworthy that some participants subjectively perceived the experience as low, even though they had played the game for a long time in terms of session data. This is also important in terms of the disconnect between subjective user experience and objective playing time. Although the participants spent time, this time did not turn into a meaningful experience, suggesting that the in-game tasks or the responses received from the artificial intelligence did not have emotional or cognitive impact.

In conclusion, these statistical findings reveal that users perceive the game system as semantically repetitive, cognitively weak, and ambiguous. The main areas for improvement include increasing the semantic diversity of the AI-generated content, clearly visualizing and intuitively communicating the prediction process to the user, and designing experiential feedback mechanisms to make the interaction sustainable and satisfying. Such improvements will increase not only the system's technical capacity but also the value it offers in terms of user experience.

Game 2 (FLAN-T5)

The user interaction times and productivity levels were analyzed separately for each difficulty level (easy, medium, hard) of the AI-assisted color guessing game in Game 2. In this game version, a pretrained large language model (LLM) was used for fine-tuning to provide users with a more controlled, meaningful, and relevant question-answer flow. In this context, the dataset includes the total time (in seconds) and the number of questions for 10 participants at easy, medium, and hard levels. The total number of question interactions with the 10 participants ranged from 170 to 240 (Table 4).

According to the statistical analysis results, the average time spent at the easy level was only 1.0 minutes (SD = 0.5), and 10.1 questions (SD = 1.7) were asked at this level. The low duration and limited number of questions at this level indicate that the model attempts to reach the goal in a shorter time with less information. At the same time, because the structure of the questions at this level was more direct and goal-oriented, the repetition rate was low, and users generally found the questions understandable (Table 5).

Table 4

Detailed Interaction Time and Question Count Across Difficulty Levels in Game 2 (FLAN-T5)

Participant ID	EASY		MEDIUM		HARD		TOTAL	
	SEC	Q COUNT	SEC	Q COUNT	SEC	Q COUNT	SEC	Q COUNT
1	139.67	13	3634.08	116	489.24	41	4262.99	170
2	56.64	9	11251.97	152	477.20	47	11785.81	208
3	51.17	10	4052.21	22	338.27	35	4441.65	178
4	54.11	9	3954.56	168	290.98	41	4299.65	218
5	48.41	9	3026.96	136	277.91	31	3353.28	176
6	72.15	13	3387.60	150	202.96	28	3662.71	191
7	52.87	10	2836.00	133	314.31	32	3203.18	175
8	44.98	8	1407.68	89	271.18	31	1723.84	128
9	49.20	10	3786.60	187	224.25	30	4060.05	227
10	52.14	10	4111.97	176	454.31	54	4618.42	240

A significant jump was observed in the frequency of medium interactions. The average duration was 69.1 minutes (SD = 43.7), and the average number of questions posed was 132.9 (SD = 48.5). This high variance suggests that the model achieved a certain stable text flow through fine-tuning; however, its productivity remained flexible according to the players' responses. Participants reported that they encountered meaningful and contextually coherent questions at the medium difficulty level and that the AI's prediction process became more visible at this level, where the sense of repetition was low. This indicates that the model successfully generated differentiated questions while maintaining the semantic context (Table 4).

The hard level was completed in an average of 5.6 minutes (SD = 1.7) with an average of 37.0 questions (SD = 8.5). At this level, according to the data reported by the participants, the model maintained focused prediction behavior, but repetition was observed. These results indicate that the model may experience semantic contraction in scenarios requiring more specificity at the difficult level; however, it is nevertheless able to maintain its productivity (Table 5).

When the overall distribution is analyzed, the interaction times and the number of questions at the intermediate level are observed to be significantly higher than at the other levels. This shows that the fine-tuned LLM model provides a relevant, meaningful, and fluid game experience, especially at the medium difficulty level. At this level, where players experienced a change in their time perception, a flow experience was provided throughout the game. When evaluated within Mihaly Csikszentmihalyi's Flow Theory, the skill-difficulty balance was established at an optimum level. Participants became more involved in the game through clear goals and constant feedback and reported losing their perception of time.

In this context, the LLM-assisted version of Game 2 demonstrated both AI-controlled content generation and high player engagement. In particular, intermediate-level data demonstrate that a system structured to optimize productivity and interaction time functions successfully.

Table 5

Summary of Average Duration and Question Count by Difficulty Level (in Minutes)

Difficulty Level	Average Duration (min)	Std. Deviation (min)	Average Question Count	Std. Deviation (count)
Easy	1.0	0.5	10.1	1.7
Medium	69.1	43.7	132.9	48.5
Hard	5.6	1.7	37.0	8.5

Model Performance and Training Findings

The training outputs obtained because of the model's fine-tuning process demonstrate that the system has successfully acquired goal-oriented production competence. The training loss and validation loss values monitored during each training epoch demonstrate that the model improved steadily in terms of semantic generalization and task fidelity.

As shown in Table 1, the training loss value, which was 3.50 at the beginning, decreased significantly within only a few epochs, reaching 0.0317 at the end of the 10th epoch. Similarly, the validation loss, which was 2.17 in the first phase, decreased to 0.0197 at the end of the training, indicating that the model moved away from a rote structure and reached the ability to produce generalizable knowledge (Table 6).

In this process, the fact that the model does not show any signs of overfitting and that the training and validation losses are close to each other indicates that the system has gained production power for training data and wider probability sets. In addition, system information about the training process (sampling rate, FLOPs value, etc.) is summarized in Table 2, and these values demonstrate the technical efficiency of the process.

Table 6

Epoch-Based Training and Validation Loss of the Fine-Tuned FLAN-T5 Model

Epoch	Training Loss	Validation Loss
1	3.5011	2.1697
2	0.6172	0.2374
3	0.1344	0.0559
4	0.0639	0.0236
5	0.0417	0.0210
6	0.0329	0.0213
7	0.0338	0.0204
8	0.0275	0.0196
9	0.0316	0.0198
10	0.0317	0.0197

These results suggest that the contextual orientation goals described in the previous section were directly reflected in the educational outcomes of the model. The model has achieved the capacity to provide grammatically correct structures and semantically diverse and contextually coherent productions. This defines the main technical achievement behind the more meaningful, less repetitive, and more strategic question productions observed in the game flow (Table 7).

Table 7

Technical Summary of the FLAN-T5 Model Fine-Tuning Process

Metric	Value
Total Training Time	10,485.35 seconds
Training Sample Rate	0.858 samples/second
Training Step Rate	0.215 steps/second
Total FLOPs (Floating Point Ops)	1.54×10^{16}
Total Training Steps	2250

Game 2 Survey Evaluation

Following the game interactions, user experience data were collected through a structured questionnaire with a 5-point Likert-type scale (n = 10).

Table 8

User Experience Survey Results for Game 2 (FLAN T-5)

Question	Number of Participants (n)	Mean (M)	Standard Deviation (SD)	Minimum (Min)	1st Quartile (Q1)	Median (Md)	3rd Quartile (Q3)	Maximum (Max)
I immediately understood the questions asked by the AI.	10	2.7	1.1	1	2	2.5	3.8	4
Similar questions were repeated repeatedly.	10	3.8	0.8	2	4	4	4	5
The AI tried guessing the color.	10	3.4	0.8	2	3	3	4	5
I lost track of time while playing the game.	10	3.5	0.7	2	3	4	4	4
The game was very engaging.	10	3.5	0.7	2	3	4	4	4

The obtained data reveal the impact of artificial intelligence (AI) on user experience using quantitative indicators and make the system's strengths and areas open to improvement visible in a concrete way. "I understood the questions asked by artificial intelligence (AI) immediately." Mean $M = 2.7$, standard deviation $SD = 1.1$, median $Md = 2.5$, minimum = 1, maximum = 4, and $n = 10$. Although this value indicates that, in general, the participants had partial difficulty in understanding the questions generated by the AI, this can be interpreted as the result of an experience in which the artificiality was not felt and a language setup that was quite close to human, rather than the failure of the system. In other words, the fact that some participants could not fully understand the questions could also be attributed to the AI's questioning style being quite realistic. Therefore, they felt that they were interacting with a real human being, not the artificial system. This result indirectly indicates that the semantic production capacity of the system based on the big language model (LLM) is strong (Table 8).

"Similar questions came repeatedly." The mean values for this statement were $M = 3.8$, $SD = 0.8$, $Md = 4$, minimum = 2, and maximum = 5. The participants' high level of agreement with this item suggests that the AI system cannot provide sufficient diversity in question production and follows an inquiry strategy based

on certain patterns. While this finding reveals that the system can be further improved regarding contextual richness, it also demonstrates that the model establishes a consistent and systematic question structure. In this context, repetitive questions can be considered not a weakness of the system but a natural outcome of the hypothesis-based prediction process (Table 8).

"Artificial Intelligence (AI) tried guessing the color." The calculated mean for this statement is *M* = 3.4, *SD* = 0.8, *Md* = 3, minimum = 2, and maximum = 5. These data reveal that participants gave a mostly positive evaluation; however, some individuals did not find the system's prediction process sufficiently visible or understandable. In this case, it is important for AI to ask questions and to be able to reflect the prediction process and logic more clearly. In the context of cognitive guidance and process tracking experienced by participants during the game, the transparency of the prediction systems was a factor that directly affected the user experience (Table 8).

"I lost track of time while playing the game," and "The game was very immersive." The means of the responses to the items were measured as *M* = 3.5, *SD* = 0.7, and *Md* = 4 for both items, respectively. The minimum value = 2, and the maximum value = 4. These findings suggest that users developed high levels of attention and interest while experiencing the game and experienced a partial shift in their time perception.

These results are significant when evaluated in the context of flow theory. The AI interaction offered by the game supports cognitive continuity by generating a constant curiosity and desire to produce answers by the user. This situation restricts the player's perception of time. The positive evaluation of the immersion level of the game shows that the system offers a satisfying structure not only technically but also on an esthetic and experiential level.

The survey results demonstrate that the LLM-based AI system has gained a meaningful and substantial place in game experience. The participants' answers to the questions demonstrate that the system is successful in terms of *naturalness* in language production, continuity in interaction, and overall integrity of the game. However, the system could be improved to offer more variety in question generation and to express the prediction process more clearly, which would further enhance the user experience. This study provides a pioneering example of how LLMs can play an active role in text generation, in-game interaction, guidance, and meaning-making processes.

Entropy and User Experience Analyses

In this study, entropy is used to evaluate the production behavior of large language models (LLMs) regarding diversity and decision-making consistency; in particular, Shannon Entropy is considered a basic measure of information theory. Developed by Claude Shannon, it allows for calculating the average amount of information carried by a message or system output and the level of uncertainty in this output (Shannon, 1948). The higher the entropy value, the more diverse and unpredictable the distribution; the lower the entropy value, the more homogeneous and repetitive the user responses to the system outputs. In this context, Shannon Entropy was used in the dynamic challenge management (DDA) process to evaluate the semantic diversity and contextual richness of the content provided by artificial intelligence systems and their impact on the player experience.

Table 9

Comparison of Shannon Entropy Values for User Experience Criteria in Two AI-Supported Game Prototypes

Metric	Game 1: Cohere	Game 2: FLAN-T5
Question Clarity	1.846	1.846
Perceived Repetition	1.295	1.357
Prediction Clarity	0.971	1.685
Flow Experience	1.157	1.361
Engagement	1.157	1.361

The analysis findings based on Shannon entropy revealed that the two different AI-powered game prototypes (Game 1 - Cohere, Game 2 - FLAN-T5) exhibited significant differences in user experience diversity. For the "Question Clarity" criterion, equal entropy values (1.846) were obtained for both games. This indicates that although the FLAN-T5 model had a higher average score, the distributional diversity was similar in both models. On the "Perceived Repetition" criterion, the FLAN-T5 model demonstrated higher entropy (1.357 vs. 1.295), indicating that some users perceived the model as repetitive and diverse by others. The most striking difference was observed in the "Prediction Clarity" metric, where the FLAN-T5 model produced a more flexible and multi-layered experience than the Cohere model (0.971), with an entropy value as high as 1.685. Similarly, FLAN-T5 showed higher entropy values in the "Flow Experience" and "Engagement" criteria (1.361 vs. 1.157). This suggests that the system offers experiences of different intensities to a broader user profile, i.e., it creates a personalized level of engagement. These findings show that Shannon Entropy is not only a statistical measure of dispersion but also a powerful analysis tool that quantitatively measures diversity in player perception and differences in experiential impact (Table 9).

Comparative Evaluation of User Experience Metrics between Game 1 (Cohere AI) and Game 2 (FLAN-T5)

The user experience comparison between Game 1 (Cohere AI) and Game 2 (FLAN-T5) reveals how the two systems differ in terms of player perception, interaction quality, and emotional connection. The proposed FLAN-T5 model exhibits significant superiority in terms of all five user experience metrics. In particular, the +2.1 point difference recorded in experiential measures such as time perception (flow experience) and immersion shows that this model offers a more balanced and satisfying structure, technically and experientially (Table 10).

The questions presented by FLAN-T5 were clearer (+0.6) and less repetitive (-0.6), thus reducing the players' cognitive load and increasing the sense of novelty in the interaction. One notable difference was that users perceived the AI's prediction behavior more clearly (+1.9). This reflects the model's success in not only asking questions but also making the prediction strategy intuitive to the player (Table 10).

In particular, the high scores obtained in the flow experience and immersion scores indicate that the FLAN-T5 model can effectively provide dynamic difficulty management (DDA) and allow the player to have experience in accordance with the skill-difficulty balance. These findings confirm the potential of controlled, diversified language output to improve production quality, interactional continuity, and player satisfaction.

This analysis supports the table and demonstrates that the relationship between output entropy and user experience is multi-layered. There is a direct link between technical variety and meaningful, guided, and satisfying gaming experiences.

Table 10

Comparative Evaluation of User Experience Criteria Between Game 1 (Cohere AI) and Game 2 (FLAN-T5)

User Experience Metric	Game 1-Cohere AI (n=10)	Game 2 - FLAN-T5 (n=10)	Difference (Δ)	Interpretation
1. Question Clarity	2.1	2.7	+0.6	FLAN-T5 is more understandable
2. Perceived Repetition	4.4	3.8	−0.6	FLAN-T5 exhibited less redundancy
3. Predictive Behavior Clarity	1.5	3.4	+1.9	FLAN-T5 is significantly more transparent than
4. Flow Experience (Time Perception)	1.4	3.5	+2.1	FLAN-T5 facilitates more substantial flow experience
5. Engagement	1.4	3.5	+2.1	FLAN-T5 is more engaging

Analysis of t-Test Results for Independent Samples

The results of the independent samples t-test demonstrate significant differences between the two AI-powered game prototypes (Game 1-Cohere AI and Game 2 - FLAN-T5) in terms of user experience metrics. In comparing five main user experience metrics, statistically significant differences were observed for four metrics, while only one metric showed results on the borderline of significance.

In the first criterion, "The comprehensibility of the questions posed by the AI", the mean scores of the user group belonging to the FLAN-T5 prototype were higher (M = 2.7), and this difference was significant at the $t(18) = 2.17$, $p = .043$ level. This result suggests that the fine-tuned model produced clearer and more comprehensible questions, thus supporting user cognition more effectively (Table 11).

The t-value for the criterion "perception of repetition of similar questions" is 5.19 and $p < .001$. This difference indicates that FLAN-T5 can produce less repetitive, diversified semantic outputs. The contextual production strategy of the model prevents users from experiencing cognitive fatigue due to repetitive patterns.

Table 11

Independent Samples t-Test Results Comparing User Experience Metrics Between Game 1 and Game 2

User Experience Metric	t-value	p-value
I immediately understood the AI questions	2.17	.043
Similar questions were repeated	5.19	< .001
The AI attempts guessing the color	−4.57	.0002
I lost track of time while playing the game.	−5.33	< .001
The game was very engaging	−5.33	< .001

The highest effect size was observed for "Visibility of the AI's prediction behavior" (t (18) = −4.57, p = .0002). FLAN-T5 users observed the system's predictive behavior more clearly, suggesting that production transparency positively affects the user experience. This finding emphasizes that not only performance but also the understandability of the process is a critical parameter in DDA (Dynamic Difficulty Adjustment) systems (Table 11).

Similarly, strong differences were observed for experiential measures such as "change in time perception" and "immersion" ($t(18) = -5.33$, $p < .001$ for both). This suggests that the FLAN-T5 model created an intense experience that enabled participants to maintain attentional focus and change their perception of time

during the game. When considered in the context of Flow Theory, these findings confirm that FLAN-T5 supports the experience of "flow" by providing a skill-challenge balance (Kaye, 2016).

In summary, these statistical findings reveal that the retrained FLAN-T5 model outperforms not only in terms of production quality but also in terms of cognitive and emotional parameters that shape user experience. It is concluded that accuracy and multi-layered criteria, such as transparency, diversity, meaningfulness, and experiential satisfaction, should be optimized simultaneously in the design of LDA systems.

Discussion

The comparative analysis of two different artificial intelligence-based game prototypes developed in this research evaluates the contributions of systems based on large language models (LLM) in the context of dynamic challenge management (DDA) on the axis of each hypothesis. The obtained data are discussed below in line with the results of user experience surveys, entropy measurements, and independent sample t-tests.

H1a: LLM-based prediction strategies enable AI systems to dynamically adjust in-game difficulty levels in real-time.

This hypothesis is supported by the semantic guidance, time-balanced interaction, and high flow scores of the FLAN-T5 model, especially at the medium difficulty level. The fact that participants reported experiencing a shift in time perception (M = 3.5) and that the system could make meaningful transitions in question generation indicates that dynamically structured difficulty adjustment was successful. In this context, hypothesis H1a was confirmed.

H1b: LLM-supported AI systems can make strategic decisions at different difficulty levels.

The data for Game 2 showed that the model could change strategy at easy, medium, and hard difficulty levels thanks to the consistency in production and clarity in prediction behavior. In particular, the high productivity and low repetition rate observed at the medium difficulty level indicate that the system can strategize based on the participants' responses. Specific and goal-oriented production at difficult levels also indicates strategic flexibility. Therefore, hypothesis H1 b was also supported.

H1c: LLM-based AI predictions have lower response times and higher decision-making flexibility than traditional machine learning algorithms.

When the prediction process of both systems was analyzedT5 tended to reach the target with fewer questions and was perceived more visibly as a "guessing agent" by the participants. In particular, while the average score of the responses to the statement "AI tried to guess the color" was M = 1.5 in the Cohere model, it was M = 3.4 in the FLAN-T5 model. This demonstrates that FLAN-T5 both produces semantically clearer questions and makes more flexible decisions in the prediction process. These findings supp

H1d: The output entropy of LLMs increases the accuracy and adaptability of AI-based prediction strategies.

The analysis results based on Shannon entropy demonstrate that the FLAN-T5 model has higher entropy values in almost every criterion, and this diversity positively affects the game experience. Especially in the "Prediction Clarity" criterion, the entropy value was calculated as 1.685 for FLAN-T5, whereas it was only 0.971

for the Cohere model. This difference reveals that entropy is a determinant not only of technical diversity but also experiential diversity. Accordingly, hypothesis H1d is confirmed.

H1e: LLM-based LDA systems provide higher consistency in game flow and player engagement than traditional methods.

This hypothesis was supported by the high mean scores (M = 3.5 for both) and statistical significance ($t(18) = -5.33$, $p < .001$) of the user group of the FLAN-T5 model on the measures of "flow" (perception of time) and "engagement" (immersion). Flow experience was measured using indicators such as keeping users' attention in the game and not realizing that time had passed. At the same time, engagement was assessed through interaction continuity and attention intensity. These findings suggest that the FLAN-T5 model offers experiential consistency through its semantic production success and its more stable and flowing interaction structure with the user. Thus, hypothesis H1e is empirically confirmed.

H0: LLM-based AI systems do not provide significant improvement in dynamic challenge management.

Hypothesis H0 is rejected considering all findings. The survey data, t-test results, and entropy analysis revealed that LLM-supported systems provide a more guided, varied, and user-centered experience.

The empirical analysis revealed that artificial intelligence systems based on large language models (LLMs) can be effectively evaluated for dynamic difficulty adjustment (DDA) strategies in digital games. The comparison of two different models, Cohere Command and FLAN-T5, provided an opportunity to comprehensively examine not only technical production differences but also key parameters that determine the player experience, such as semantic cohesion, interaction quality, and flow continuity.

The primary user experience differences between the two prototypes are summarized in the table below. The comparative data in the table present directly observable results in line with the findings detailed in the discussion section (Table 1).

Table 12

Comparison of user experiences

Game 1 (Cohere AI) (n=10)	Game 2 (FLAN-T5) (n=10)
Questions were less clear; participants had difficulty understanding them.	The questions were clearer and easier to understand.
High sense of repetition; questions were very similar.	Less repetition; questions were more varied.
The prediction behavior was unclear; most participants did not notice it.	The prediction process was more visible; most participants noticed it.
The game was not fluent; participants did not lose track of time.	A flow experience occurred; participants lost track of time.
The game was less engaging; interaction felt limited.	The game was more engaging; it provided an attention-grabbing experience.

The Cohere Command model is based on zero-shot prompting; thus, its production behavior is limited. Entropy analyses show that this model offers low diversity, especially in terms of the visibility of prediction behavior (Prediction Clarity Entropy: 0.971) and high question repetition (Perceived Repetition Entropy: 1.295). In addition, the Flow experience performed poorly (Flow Entropy: 1.157), indicating that the model's interaction with the players was limited and far from satisfying. In some sessions, it was observed that

content that was disconnected from meaning, lacked contextual integrity, and was hallucinatory was generated; this negatively affected the players' decision-making processes and increased their cognitive load. The survey findings also support these results: Participants reported low clarity (M = 2.1), high repetition (M = 4.4), and poor flow (M = 1.4). It appears that a high production volume does not guarantee content quality but rather negatively affects the quality of interaction. In addition, users reported that they were often unable to notice the AI's predictive behavior, indicating a lack of transparency in the interaction process (see Table 1, row 3).

On the other hand, the FLAN-T5 model, customized by fine-tuning, produced more balanced, contextualized, and guided questions. In particular, the entropy value obtained at the medium difficulty level (0.489) indicates that the model produced meaningful, diverse, and semantically coherent questions. Participants perceived the prediction process as more open (M = 3.4), experienced a shift in time perception (M = 3.5), and found the game immersive (M = 3.5), suggesting that this model offered a stronger structure in terms of experience. These data suggest that users rated FLAN-T5 as a more understandable, fluid, and interactive system (see Table 1, rows 1 and 4).

Particularly noteworthy is that meaningful but not directly predictive questions in the FLAN-T5 game prototype (e.g., "Does this color occur frequently in nature?") maintained player focus and prevented hallucinatory content. By smoothing the transition between difficulty levels, such questions balanced the player's cognitive load and ensured a more controlled DDA process. This structure can be described as a "difficulty buffer" in the literature.

Moreover, the more explicit and intuitive realization of the prediction behavior in the FLAN-T5 model positively affected the game functioning and the strategic relationship that the player established with the system. This suggests that DBDA systems should be designed not only at the algorithmic level but also in a way compatible with player perception. On the other hand, in the Cohere Command model, the uncertainty of the prediction process, the inability to feel the difficulty transitions, and the low flow experience reveal that the system's lack of transparency weakens the game experience.

As a result, LLM strategies make sense not only in terms of the amount of content production but also in terms of semantic quality, contextual coherence, and player intuitiveness of the production process. In this context, the following conditions must be satisfied for LLM-based systems to provide a successful LDA experience:

1. Generate meaningful content with a balanced level of entropy.
2. Strategic management of contextual transitions
3. Balancing production processes with meaningful intermediate content
4. The prediction process is presented to the player clearly and intuitively.

Although the FLAN-T5 model met these conditions to a large extent and provided a more experientially successful structure, the Cohere Command model failed to provide interaction integrity despite its production volume. The research provides a powerful model of how LLM-based AI systems can be optimized in DDA processes with technical accuracy, semantic routing, flow continuity, and controlled entropy management centered on player experience (Table 12).

These findings are consistent with the theoretical framework of the study. Participants' feedback, such as shifts in time perception, concentration of attention, and feelings of productivity, fulfills the ideal experience conditions described in Csikszentmihalyi's (1990) Flow Theory (Cowley et al., 2008). In particular, the fact that

the FLAN-T5 model posed meaningful and goal-oriented questions to the player supported the continuity of the flow state by balancing difficulty levels with skill levels. However, when evaluated within the framework of Sweller's (1988) Cognitive Load Theory, repetitive, semantically incomplete, or hallucinatory productions reduced the quality of experience and caused distraction in the Cohere model. Moreover, when evaluated within Ryan and Deci's (2000) Self-Determination Theory, the FLAN-T5 model reinforced intrinsic motivation by offering players more autonomy, competence, and interaction. Overall, large language models can be repositioned not only as producers of technical output but also as dynamic actors providing semantic guidance, experience management, and flow continuity. However, besides these strengths, some practical limitations in integrating LLMs into game systems should also be considered. In particular, most LLMs work with credit-based systems through API access, and each user interaction incurs a certain cost. This makes economic sustainability difficult in highly interactive game scenarios and can be limiting, especially for independent developers and research-oriented projects. Furthermore, the dependency of LLM systems on technical components such as server response time, connection stability, and data security can lead to problems such as delays or response inconsistencies in real-time gameplay. For these reasons, the large-scale integration of LLM-based systems into games requires optimization of not only model performance but also economic accessibility and technical compatibility.

Conclusions and Recommendations

By comparing the game-based production performance of the zero-shot prompting Cohere Command model and the semantically fine-tuned FLAN-T5 model, this research demonstrates that large language models (LLMs) should be evaluated not only in terms of output accuracy but also in terms of contextual directiveness, semantic variety, and experiential impact. The findings demonstrate that the FLAN-T5 model, in particular, makes the interaction more sustainable, flowing, and satisfying by generating more meaningful and goal-oriented questions, reducing repetition, and making the prediction process more explicit and intuitive to the player. Accordingly, it is suggested that artificial intelligence, especially in narrative-driven or role-playing games (RPGs), should not be limited to text generation but should be integrated with high-level cognitive processes such as meaning-making, contextual guidance, prediction strategy development, and interaction continuity. Transparently perceivable prediction behavior of the AI by the player plays a critical role in system reliability and quality of experience. When evaluated in the context of Flow theory, production formats that provide flow, especially at medium difficulty levels, support high user engagement by balancing the experience. In this framework, game systems should be developed with technical success and structures that can immerse the user on cognitive and emotional levels. Future studies suggest that scenarios should be developed enriched with sensory inputs such as voice commands, gesture recognition, and emotional state analysis, where LLM-based artificial intelligence systems can be integrated with multi-modal data inputs. Integrating such systems into in-game narrative structures and dialog systems will enable more human, intuitive, and seamless interaction between AI and users. Furthermore, the design of visual interfaces that can intuitively convey the decision logic of artificial intelligence to the player, mechanisms that test player perception with gamified experimental scenarios, and adaptive system architectures that manage the production process with semantic filters will deepen the research in this field at both technical and experiential levels. In this context, this study fills an important theoretical and practical gap by opening up the behavioral and experiential performance of different LLM architectures in game-based DDA scenarios, which have not yet been systematically compared in the literature. The study strongly suggests that the

role of AI in digital game design is not only as a generative tool, but also as an active and redefined design component in the construction of meaning, guidance, and experience.



Ethics Committee Approval: Ethics committee approval was received for this study from the ethics committee of Doğuş University (Date: 28.05.2025, No: 82320).
Informed Consent: Written informed consent was obtained from all participants who participated in this study.
Peer Review: Externally peer-reviewed.
Conflict of Interest: The author has no conflict of interest to declare.
Grant Support: The author declared that this study has received no financial support.

Author Details
Onur Aşkın
[1] Doğuş University, Faculty of Art and Design, Department of Digital Game Design, İstanbul, Türkiye
0000-0002-1928-474X oaskin@dogus.edu.tr

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.* https://doi.org/10.48550/arXiv.2005.14165

Cameirão, M. S., Bermúdez i Badia, S., Duarte, E., & Verschure, P. F. M. J. (2009). *The Rehabilitation Gaming System: A review. Studies in Health Technology and Informatics, 145*, 65–83. https://doi.org/10.3233/978-1-60750-018-6-65

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, 50(4), 31-34. https://doi.org/10.1145/1232743.1232769

Colwell, A. M., & Glavin, F. G. (2018). Colwell's castle defence: a custom game using dynamic difficulty adjustment to increase player enjoyment. *arXiv preprint arXiv:1806.04471.*

Cowley, B., Charles, D., Black, M., & Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment (CIE)*, 6(2), 1-27. https://doi.org/10.1145/1371216.1371223

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper and Row.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. Basic Books.

Csikszentmihalyi, M. (2014). *Applications of flow in human life*. Springer.

Fisher, N., & Kulshreshth, A. (2024). Exploring dynamic difficulty adjustment methods for video games. *Virtual Worlds*, 3(2), 230-255. https://doi.org/10.3390/virtualworlds3020012

Garcia-Ruiz, M., Montesinos-López, O. A., & Anido-Rifón, L. E. (2023). The use of deep learning to improve player engagement in a video game through dynamic difficulty adjustment based on skills classification. *Applied Sciences*, 13(14), 8249. https://doi.org/10.3390/app13148249

Gilleade, K. M., Dix, A., & Allanson, J. (2005, January). Affective videogames and modes of affective gaming: assist me, challenge me, emote me. In *Proceedings of DiGRA 2005 Conference: Changing Views: Worlds in Play*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Kaye, L. K. (2016). Exploring flow experiences in cooperative digital gaming contexts. *Computers in Human Behavior*, *55*, 286-291. https://doi.org/10.1016/j.chb.2015.09.023

Keller, J., & Bless, H. (2008). Flow and regulatory compatibility: An experimental approach to the flow model of intrinsic motivation. *Personality and Social Psychology Bulletin*, 34(2), 196-209. https://doi.org/10.1177/0146167207310026

Kowlessar, T. (2020). *How Difficulty Affects Player Engagement in Digital Games* (Doctoral dissertation, Flinders University, College of Science and Engineering.).

Lopes, R., & Bidarra, R. (2011). Adaptivity challenges in games and simulations: A survey. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(2), 85-99. https://doi.org/10.1109/TCIAIG.2011.2152841

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.

Roohi, S., Guckelsberger, C., Relas, A., Heiskanen, H., & Takatalo, J. (2021). Predicting game difficulty and engagement using AI players. *arXiv preprint arXiv:2107.12061.* https://doi.org/10.48550/arXiv.2107.12061

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68-78. https://doi.org/10.1037/0003-066X.55.1.68

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3), 1-24. https://doi.org/10.1145/1077246.1077253

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008). https://doi.org/10.48550/arXiv.1706.03762

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261-272. https://doi.org/10.1038/s41592-020-0772-5

Wu, C.-H., Chen, Y.-S., & Chen, T.-C. (2017). *An adaptive e-learning system for enhancing learning performance: Based on dynamic scaffolding theory*. *Eurasia Journal of Mathematics, Science and Technology Education, 14*(3). https://doi.org/10.12973/ejmste/81061

Yannakakis, G. N., & Togelius, J. (2018). *Artificial intelligence and games*. Springer. https://doi.org/10.1007/978-3-319-63519-4

Zheng, T. (2024). Dynamic difficulty adjustment using deep reinforcement learning: A review. *Applied and Computational Engineering*, *71*, 157-162. https://doi.org/10.54254/2755-2721/71/20241633