Prompted: a puzzle game on LLM prompting

Gijs van der Heide, Jordy del Castilho, Leonardo Lago, Pavel Verigo, Teodor Stănilă, Zeyu Fan, Tzu-Hsin Hsieh, Amir Zaidi
Delft University of Technology
Delft, Netherlands

Giulio Barbero, Mike Preuss
Leiden University
Leiden, Netherlands

Rafael Bidarra
Delft University of Technology
Delft, Netherlands
R.Bidarra@tudelft.nl

***Abstract*—Large Language Models (LLMs) have become increasingly widespread in recent years. The quality of an LLM's output heavily depends on the prompt it receives. Therefore, users giving a task to an LLM should not only understand that task's purpose and constraints, but also clearly communicate them to the LLM. Although several games use prompting as a central mechanic, there is no clear example of a game designed specifically to help players practice their prompting skills. We present *Prompted*, a web-based puzzle game in which players solve a series of themed levels by interacting with an LLM. The objective is to get an LLM's output to satisfy the level's requirements, while constraints are placed on the players' prompts. Preliminary playtest of *Prompted* was very positive. However, a thorough evaluation of the game's efficacy has still to take place.**

***Index Terms*—LLM, Prompting, Serious games, Puzzle-based learning**

I. Introduction

In recent years, Large Language Models (LLMs) have become commonplace in most people's lives; from education to professional environments, they are frequently used both in menial and more complex tasks. However, LLMs' effectiveness varies depending on the context of application and the users' experience in writing appropriate prompts. Although AI models seemingly get smarter by the day, the way a request is formulated still plays an important factor in producing the expected output. The risk of hallucinations remains a problem, and getting unexpected responses due to imprecise prompting can add unnecessary frustration to the users' experience. Therefore, it is worth studying how to improve prompting skills and effectiveness.

Currently, LLM-related training focuses on responsible usage of LLMs. This includes knowledge about functioning, accuracy and ethical concerns [1]. Very little has been developed specifically to target prompting skills. Moreover, these skills are usually developed with practice and long tedious sessions of trial and error. A game could potentially facilitate such practice while fostering engagement, raising the question: *"How effective can a video game be as an environment to practice and improve prompting skills?"*

To answer this question, we designed and developed the serious game *Prompted*. *Prompted* is a web-based puzzle game that lets players interact with LLMs while simultaneously stimulating improvement of their prompting skills with increasingly complex tasks. In playtesting sessions, we found that players experienced *Prompted* as intuitive and easy to play, and self-reported as feeling more confident in their prompting abilities after playing.

II. Related Work

We briefly survey previous work on puzzle-based learning, prompting-related games and the development of prompting skills.

A. Puzzle-Based Learning

Educational research shows that game-based learning, especially puzzle games, can significantly boost engagement, motivation and critical thinking [2]. Serious games can turn passive learners into active ones by embedding educational content in interactive challenges. For example, Khorammakan et al. find that puzzle game–based learning yields large gains in knowledge, cognitive performance, and problem-solving skills [3]. Similarly, Schrader reports that serious games strongly affect cognitive outcomes and learner motivation [4]. Key benefits of puzzle-based learning include immediate feedback and incremental difficulty, which reinforce learning. This type of challenge encourages players to experiment with different solutions, take risks, and learn from mistakes in a low-stakes setting.

The specific advantages that lead us to choose video games as a medium for skills development are the following:

- Active problem-solving: puzzle games engage learners in solving well-defined challenges, improving critical thinking and strategy.
- Incremental complexity: levels and clues scaffold learning; success at one stage builds confidence for the next.
- Immediate feedback: players instantly see the consequences of their actions (e.g. correct/incorrect puzzle inputs), enabling rapid trial-and-error learning.
- Engagement and motivation: a compelling narrative and gameplay (rather than lectures) keeps learners curious and invested in the content, motivating continued play and deeper learning.

B. Prompt-as-Puzzle Games

Several recent games illustrate this approach. For instance, *Prompt Override* [5] is a hacking-themed game where players

explore a simulated file system and edit an LLM's system prompt to solve puzzles. The authors explain that by having the ability to directly manipulate system prompts in their game, solutions become open-ended and players develop a deeper understanding of prompt engineering. Another example is *Cracking Aegis* [6], a dialogue-based adventure which makes users impersonate characters and issue prompts in order to extract sensitive information. The game flow is essentially practicing harder and harder prompting techniques in order to succeed. After playing, participants report better awareness of issues like phishing and AI misuse, showing the pedagogical impact of the prompt-driven gameplay. These examples suggest the viability of *Prompted*, as both games use prompting as the primary means of puzzle-solving, directly aligning with its practicing goal. However, in contrast to those games, *Prompted* uses a variety of puzzle types, each involving different goals and techniques related to working with LLMs. Player prompts may be subject to constraints, such as a character or word limit, and a global leaderboard adds pressure and motivation to improve.

C. Prompting Skills

Recent studies emphasize that prompt engineering, the skill of writing effective inputs for AI models, is both important and learnable. Lee and Palmer [7] note that prompt engineering "can be learned and developed" and thus should be explicitly taught. However, pedagogy for this skill is still emerging [8]. Well-designed prompts can greatly improve AI outputs, making prompt engineering a valuable skill for students to develop, not as a replacement for critical thinking, but as a way to work more efficiently with AI. The growing number of prompt engineering courses and resources, for example, Arizona State's AI prompt course [9], reflects a recognized need to incorporate it into curricula. By framing prompt engineering as an active, strategic process, educators can use interactive methods to teach it.

A puzzle-game format is well-suited to teaching these skills because it mirrors the trial-and-error nature of crafting prompts [10]. Prompt engineering often involves testing different instructions and adapting based on results, much like solving a puzzle. This type of game naturally provides clear goals with constraints, iterative refinement, and skill progression.

By carefully integrating these elements, a game could guide users to discover effective prompting strategies organically. Then, as the levels progress, players could practice framing prompts of increasing sophistication and learn to reason about how phrasing affects the AI's output. In short, we can conclude that puzzle mechanics promote the same analytical and creative thinking needed for prompt engineering.

III. Game Design

We describe *Prompted*'s game mechanics, its game loop and main design features, including the scoring mechanism, as well as the game look and feel.

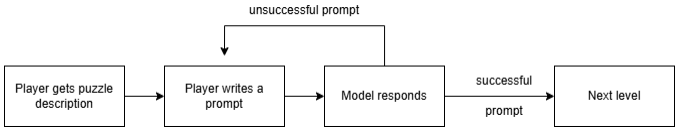

Fig. 1. *Prompted* Game Loop

A. Synopsis

Prompted takes the player through a series of increasingly harder levels that incorporate different aspects and capabilities of LLMs, such as function calling and multi-modal perception. The game is structured around a repeating loop that mirrors the real world process of iterative prompt refinement (see Figure 1). Players have to improve their prompt to be more specific or think of different techniques to be able to reach the required outcome.

All of the challenges can be accessed through a world map featuring a main path for game succession, as well as branching paths to more innovative or experimental levels, for example multiplayer challenges (see Figure 2). The multiplayer levels pair two players and make them work towards the same goal, getting the AI to print the same output. The challenge can be transparent, allowing for all prompts and outputs to be visible, but also opaque, where each players has access only to their AI conversations.

The puzzles were designed with one or more of the following purposes:

1) Increasing players' awareness of the fact that a prompting strategy can significantly affect the output quality of an LLM.
2) Helping players become more effective prompters, and achieve desirable outputs more efficiently.
3) Teaching players about the limitations of LLMs.

A skip button becomes accessible after five attempts for each stage so that players do not get stuck on a level they cannot complete yet. After ten failed attempts, an extra hint is presented. Upon completing a level, the player's score is added to the leaderboard allowing players to compare their performance.

The level screen presents the task description, which is revealed progressively through clicking the game mascot and the description box alongside a text box for the player's prompt (see Figure 3). Buttons for the map, the leaderboard, and the skip feature, if unlocked, are also available. Additionally, it displays the player's previous prompts for the level and the model responses.

B. Game Mechanics & Style

Among the essential game mechanics, we should point out:

1) Continuous feedback loop, to help players refine their prompt and learn about why their previous attempts would not lead to the desired output.
2) Various LLM concepts, such as token limits and hallucination, woven into the level design, ensuring the game is educational without sacrificing fun.

3) A leaderboard where players can see how effective their prompt is compared to others, to keep the game competitive and reward effective prompts.

In addition, we introduced several other secondary elements aimed at helping achieve the game's purpose. These include:

1) A tutorial that helps players understand the main concept of *Prompted*, onboarding them regarding level setup and the actions available.
2) A defined number of stages, providing a good game overview and keeping it accessible.
3) A profile page showing all scores of the player's passed levels and the player's settings.
4) A global game map, allowing players to take different branching paths, for example, when they get stuck or want to try a different level. It also makes the game feel more engaging, as users are able to track their progress throughout the map in a colorful way.
5) A mascot character to make the game have more of its own identity.

At the top level, *Prompted* presents a map overview that allows players to select and enter levels, as depicted in Figure 2. Each level consists of the puzzle description, where the desired output is requested, an input text box for prompting, and a stack showing the output of each past prompting attempt, as depicted in Figure 3. The visuals are designed to be attractive but minimalist in order not to distract and to emphasize functionality.

C. Scoring metrics

To calculate the scores on the global leaderboard, we consider only the length of the winning prompt. Shorter prompts yield higher scores, as they reflect more efficient and concise interactions with the LLM.

At the same time, the number of attempts used to pass each level is also recorded and displayed on the leaderboard, though it does not affect the final score. This preserves transparency and still encourages players to solve levels in fewer attempts.

Prompted differentiates between the first score and the best score. The *first score* is obtained the first time a player passes a level and depends on the length of the prompt used at that moment. It cannot be changed afterwards.

Fig. 2. *Prompted* Level Map

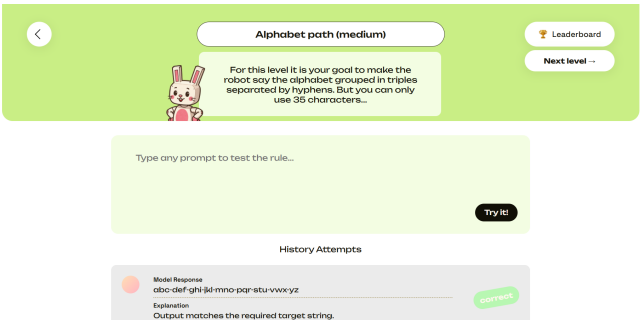

Fig. 3. Example of a Level Screen

Fig. 4. High-level system overview.

The *best score*, on the other hand, reflects the shortest prompt ever submitted by that player for that level, regardless of how many attempts were made. These scoring methods motivate users both to initially make concise prompts and to revisit levels to further refine their prompting strategies.

For levels that incorporate editing images, we use two similarity metrics to determine how close players got to the target: the root mean square error (RMSE) [11] and the structural similarity index (SSIM) [12]. RMSE measures the differences in pixel values between the generated and reference images, while SSIM uses luminance, contrast and structural features.

IV. Implementation

Prompted is a web-based game; all the player has to do is connect to the URL to start it[1] (after a one-time player name creation). This design choice makes it so people are not tied to a specific device, and the game is easy to update, since new puzzles can simply be added to the map.

On a high level, the game is a server that provides a web page for each puzzle and forwards player prompts to an API to generate an LLM reply. As shown in Figure 4, the browser sends the prompt to the server, the server adds level context, forwards it to OpenRouter and the LLM reply is then returned to the page.

Regarding the system architecture, we can distinguish the following modules:

1) Backend: The backend is written in **Python**. It uses **AIOHTTP** to set up endpoints, WebSockets, and serve HTML frontend files to players. Other options have been considered (such as Flask), but they do not, for example, have out-of-the-box WebSockets support or are simply harder to use.

2) Frontend: **JavaScript** is the natural choice for the frontend of a web-based game. No advanced frameworks are used

[1]https://api.prompted.fit/

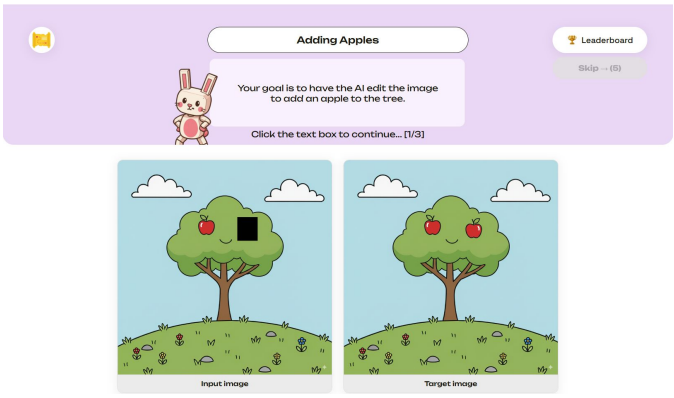

Fig. 5. Example of an image-processing task in *Prompted*

to prevent unexpected breaking changes in the future caused by third party libraries.

3) API: Using local models would significantly limit the possibilities for this game. They are less "intelligent" and/or a lot slower than large cloud models, especially considering the game should be able to run on mobile devices. *Prompted* uses LLMs through the **OpenRouter API** in order to increase the number of available models.

4) Text model: Prompted's text model is currently **Gemini 2.5 Flash**. This model is the lightweight version of Gemini 2.5, has a large enough context, and works well with tools and image recognition. This model is also relatively inexpensive compared to similar others. It is also possible to switch to an even cheaper model, but these might not be sufficient to solve some of our puzzles.

5) Image edit model: The image model is currently **Google Nano Banana** [13], [14]. This model is different from normal image generation models as it can change parts of images without changing the rest of the image, which opens up interesting possibilities for puzzle design. *Prompted* uses this model for AI-based image-processing tasks; see Figure 5 for an example.

6) Miscellaneous technologies: **Langchain** helps with abstractions and prompt management. It makes it easier to include tools and add images to prompts.

V. Conclusion

Although LLMs are gaining widespread application, many people are still using them at a very basic level. Helping such users better realize how LLMs actually behave can strongly improve their prompting skills and, therefore, their performance and satisfaction.

The game *Prompted* was designed to facilitate players practicing LLM prompting. From preliminary playtesting, its effectiveness was apparent, clearly increasing player awareness of the fact that their prompting strategy matters, reportedly making them more capable prompters after playing the game.

The choice for a web-based puzzle game appears well-founded: players reported it as easy to learn and flexible, with moderate to high perceived gains in prompting ability. Objectively, the game also seems intuitive, with just under 90% of entered prompts in play testing indicating players clearly understood the goal of the level.

Future iterations of *Prompted* could be built upon these findings. For example, the game could detect whether a player includes certain structural cues in their prompt and provide immediate feedback or small rewards when advanced prompting patterns are used. Levels could also be progressively designed to require specific strategies —such as explicitly assigning a role, decomposing the task, or constraining the output format—, so that players gradually internalize a more deliberate and modular approach to prompt design. Such an adaptive feedback system would help transition players from intuitive prompting to conscious, expert-level formulation, reinforcing the game's educational value while keeping its playful nature.

References

[1] A. Almufarreh, A. Ahmad, M. Arshad, C. W. Onn, and R. Elechi, "Ethical implications of chatgpt and other large language models in academia," *Frontiers in Artificial Intelligence*, vol. 8, 2025.

[2] Z. Yu, M. Gao, and L. Wang, "The effect of educational games on learning outcomes, student motivation, engagement and satisfaction," *Journal of Educational Computing Research*, vol. 59, no. 3, pp. 522–546, jun 2021. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0735633120969214

[3] R. Khorammakan, R. Khajouei, and R. Farrahi, "Puzzle game-based learning: a new approach to promote learning of principles of coronary artery bypass graft surgery," *Education and Information Technologies*, vol. 23, no. 241, 2023.

[4] C. Schrader, *Serious Games and Game-Based Learning*. Springer Nature Singapore, 2023, pp. 1255–1268.

[5] R. Gallotta, A. Liapis, and G. Yannakakis, "Prompt Override: LLM hacking as serious game," in *IEEE Conference on Games (CoG)*, 08 2025, pp. 1–2.

[6] Z. Fu, Y. Lin, and H. Chen, "Cracking aegis: An adversarial llm-based game for raising awareness of vulnerabilities in privacy protection," in *Proceedings of the IEEE Conference on Games (CoG '25)*, 2025. [Online]. Available: https://arxiv.org/abs/2505.16954

[7] T. Lee and A. Palmer, "Prompt engineering in higher education: a systematic review to help inform curricula," *Journal of Applied Learning and Teaching*, 2025. [Online]. Available: https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-025-00503-7

[8] Y. Qian, "Prompt engineering in education: A systematic review of approaches and educational applications," *Journal of Educational Computing Research*, 2025. [Online]. Available: https://journals.sagepub.com/doi/10.1177/07356331251365189

[9] A. Maynard, "An introduction to basic prompt engineering with ChatGPT," 2023, arizona State University. [Online]. Available: https://andrewmaynard.net/an-introduction-to-basic-prompt-engineering-with-chatgpt/

[10] I. Granic, A. Lobel, and R. C. M. E. Engels, "The benefits of playing video games," *American Psychologist*, vol. 69, no. 1, pp. 66–78, 2014. [Online]. Available: doi:10.1037/a0034857

[11] ScienceDirect, "Root mean square error," 2026. [Online]. Available: https://www.sciencedirect.com/topics/engineering/root-mean-square-error

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[13] Google, "Introducing Gemini 2.5 Flash Image, our state-of-the-art image generation and editing model," https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025, accessed: 6 Nov 2025.

[14] ——, "Nano Banana AI image generation," https://nanobananaai.ai/, 2025, accessed: 6 Nov 2025.