Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Promptémon: Assessing a Game-Based Learning Tool for Teaching Prompt Engineering to Middle Schoolers

**Priyanka Kumar Priyanka**

kumar_p@utpb.edu

The University of Texas of the Permian Basin

**Panhapiseth Lim Lim**

The University of Texas of the Permian Basin

**Kalpathi Subramanian Kalpathi**

The University of North Carolina at Charlotte

**Article**

**Keywords:**

**Posted Date:** November 23rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11368129/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Promptémon: Assessing a Game-Based Learning Tool for Teaching Prompt Engineering to Middle Schoolers

Panhapiseth Lim[1], Priyanka Kumar[1*], Kalpathi Subramanian[2]

[1,1*]Department of Computer Science,
The University of Texas Permian Basin, Odessa, TX, 79762, USA.
[2*]Department of Computer Science,
The University of North Carolina at Charlotte, Charlotte, NC, 28223, USA.

*Corresponding author(s). E-mail(s): kumar_p@utpb.edu;
Contributing authors: lim_p65274@utpb.edu; kumar_p@utpb.edu;
krs@charlotte.edu;

**Abstract**

With the increasing prevalence of artificial intelligence (AI) in daily life, it is essential for young people to be proficient in communicating with AI systems. This research paper introduces a game-based tool named Promptémon for middle schoolers, designed to teach prompt engineering: phrasing instructions so an AI gives a useful answer. It uses a creature-battle game format where players create prompts to defeat AI opponents. Players combine prepared instructions with "modifier chips" that represent personas, styles, or audiences. The game gives immediate feedback, using a second AI step to check whether the response meets the challenge. The research questions were: (RQ1) How do middle school students' confidence and self-reported understanding of prompt engineering compare before and after playing? and (RQ2) How do students respond to the game's features for trying and refining prompts? Promptémon was tested with students in grades 6 to 8 (60 forms before, 60 after) at a public school AI event in a Texas city, in visits of about seven to ten minutes. Self-reported confidence (4 or 5 of 5) rose from 48.3% to 78.3% (29 to 47 of 60). Of 53 matched students, 26 rated themselves higher, 8 lower, and 19 the same (sign test, Holm-adjusted $\boldsymbol{p = .0059}$). After play, 44 of 60 agreed the feedback helped them understand how to improve, while 28 of 60 wanted to try other chip combinations. However, choice of the most structured prompt did not rise (31, then 27 of 52; McNemar

test, $p = .22$), and only 5 of 60 named a prompt component in their own words. Because all outcomes are self-reports from one short session with no comparison group, these results describe immediate reactions, not gains in prompting skill. This study presents a game-based way to show young learners the parts of a prompt.

**Keywords:** Prompt engineering, AI literacy, Educational games, Middle school, Large language models, Game-based learning

**Short title (running head):** Assessing a game-based tool for teaching prompt engineering

**Corresponding author:** Priyanka Kumar, Department of Computer Science, The University of Texas Permian Basin, Odessa, TX 79762, USA. Correspondence to: kumar_p@utpb.edu

# 1 Introduction

Large language models (LLMs) such as ChatGPT and Google Gemini are now part of how children search for information, complete homework, and play. These artificial intelligence (AI) systems write text in reply to typed instructions, called prompts. Effective use of these systems depends on *prompt engineering*: the practice of carefully phrasing input instructions so that the model produces useful output (White et al. 2023). Much published prompt-engineering instruction, however, targets undergraduates or adults (Xiao et al. 2025; Ma et al. 2025). Younger learners, who already use LLM-backed tools, have few age-appropriate resources. Teaching middle schoolers prompt engineering is hard because the relevant concepts (audience, persona, and register, meaning how formal the language is) are abstract and model behaviour is unpredictable. Unguided trial-and-error also leads to uneven progress (Woo et al. 2025).

Prompting is hard to learn even for adults. In a study of people without AI training, Zamfirescu-Pereira et al. (2023) found that participants explored prompts opportunistically, trying whatever came to mind, rather than systematically. They drew general rules from single successes or failures and expected the model to follow instructions the way a person would. A learner who sees one prompt "work" has little reason to ask which part of it made the difference. One way to help is to make the parts of a prompt visible, so that a learner can change one part at a time and see what changes in the output. Promptémon is built on that idea.

We present a browser-based game, Promptémon, in which middle schoolers assemble prompts from prepared parts to battle AI opponents. The design borrows the creature-battle structure of games like Pokémon: players choose a starter, then attack opponents by composing prompts from a base instruction and three categories of modifier "chips" (persona, style, and audience). A persona chip sets the role the model plays, a style chip sets how it writes, and an audience chip sets who it writes for. A first call to the Gemini model generates a response; a second call asks the same model

whether the response meets the opponent's challenge, and a success defeats the opponent. The three chip categories are not arbitrary. They map onto components named in education-oriented prompting frameworks, including the PARTS components *Persona*, *Theme*, and *Recipients* (Park and Choo 2025) and Lo's CLEAR framework's emphasis on adapting a prompt and reflecting on its output (Lo 2023).

Students in Promptémon do not type free text. They choose one prepared base prompt and at most one chip in each category. The game is therefore designed to teach a narrow part of prompt engineering, namely that a request has parts and that changing one part changes the output. Writing a complete prompt for a new task is a different skill, and this study does not measure it. This limit shapes our claims and the research questions below.

We evaluated Promptémon with middle school students at one school event, collecting 60 pre-game and 60 post-game forms, from which 53 students could be matched. Our contributions are the design of the tool, an exploratory survey evaluation of one short session, and a discussion of how the design choices may relate to students' responses.

Two research questions guided the study. **(RQ1)** How do middle school students' confidence and self-reported understanding of prompt engineering compare before and after a short Promptémon session? We examine (a) self-reported confidence in getting useful AI responses, (b) agreement that persona, style, and audience can change AI output, (c) choice of prompting strategy before and after play, and (d) how students describe a good prompt in their own words. **(RQ2)** How do students respond to the game's mechanics for experimenting with and refining their prompts? We examine (a) self-reported motivation to try different chip combinations, (b) perceived helpfulness of the in-game feedback, and (c) what students said about the chip system and about learning through a game.

# 2 Related work

## 2.1 AI concepts for K-12 learners

Touretzky et al.'s "Five Big Ideas in AI" framework (Touretzky et al. 2019) and the AI4K12 initiative (AI4K12 2020) provide grade-band guidance for K-12 AI instruction (K-12 is US school education from kindergarten to grade 12). Promptémon relates most closely to the big idea that AI4K12 calls "natural interaction". That idea concerns the knowledge an AI system needs to interact naturally with people. The game looks at the same interaction from the person's side and asks how a learner instructs the system. Reviews of AI literacy have covered school programmes (Ng et al. 2024) and, separately, higher and adult education (Laupichler et al. 2022). In K-12 education, recent work on prompt engineering has mostly studied prompts that researchers and teachers write to run educational tools, rather than teaching students to prompt (Chen et al. 2024; Li 2024; Arrington et al. 2025). This gap motivated our choice to build for ages 11–14. Long and Magerko (2020) define AI literacy as a set of competencies that includes communicating and collaborating effectively with AI and using it as a tool. We put that view into practice by making prompt construction the core gameplay. Lee et al.'s middle school AI curriculum (Lee et al. 2021) and Ali et al.'s constructionist

curricula for primary and middle schoolers (Ali et al. 2019) both combine hands-on building with discussion of AI concepts. Promptémon adopts that orientation: every battle requires students to assemble a prompt before any AI response is produced.

## 2.2 Teaching prompt engineering

A small but growing literature treats prompt engineering as a teachable skill. Park and Choo's practical strategies for educators include the PARTS components (Persona, Aim, Recipients, Theme, Structure) (Park and Choo 2025). We adopted three of these components (Persona, Theme/Style, and Recipients/Audience) as our chip categories because each one can be varied on its own and each changes the output in a way a student can see. The opponent's challenge fixes the Aim, and the base prompt fixes the Structure. White et al.'s prompt-pattern catalogue (White et al. 2023) includes patterns for refining a prompt across turns, and Lo's CLEAR framework (Concise, Logical, Explicit, Adaptive, Reflective) asks users to adapt prompts and reflect on outputs (Lo 2023). This shaped our decision to allow several attempts per battle. Studies of how learners prompt show why structure helps. Woo et al. (2025) found that secondary students using ChatGPT for a writing task learned to guide it by trial-and-error on their own, some in more advanced ways than others; the authors argued for explicit instruction. Trial-and-error becomes problematic when feedback is ambiguous or absent, because the student treats whatever "works" as the rule and stops searching for better strategies. Adults without AI training show the same pattern (Zamfirescu-Pereira et al. 2023). Tolzin et al. propose worked examples of high-quality prompts as a way to build prompting skill, and their first findings link prompting skill to the quality of model output (Tolzin et al. 2024). This informed our use of pre-built base prompts and chip choices rather than free-text input. Teaching students how to prompt has been studied with undergraduates (Xiao et al. 2025) and argued for in education more broadly (Walter 2024), but it has rarely been studied with younger learners.

Two lines of work come closest to Promptémon. Ali et al. (2024) worked with teachers to design K-12 learning materials on text-to-image generation. PicPrompt, an earlier game from our group, turned prompting into a multiplayer guessing game in which one student writes a prompt for an image and the others guess it (Ngo et al. 2025; Kumar et al. 2026). ImaginAItion, a game for generative AI literacy, targets adults without deep knowledge of the technology (Ma et al. 2025). Promptémon differs from PicPrompt in three ways: its prompts produce text rather than images, it names the prompt components and makes them selectable, and an AI judge rather than other players decides the outcome. These choices let one student play alone in a few minutes, but they give up the free-text writing and the peer feedback that PicPrompt provides. To our knowledge, no prior work has used gameplay built around these components to teach middle schoolers text prompting.

## 2.3 Existing game-based AI tools for learning

Self-determination theory predicts that games support engagement when they satisfy autonomy, competence, and relatedness (Ryan et al. 2006). Promptémon is a short

single-player activity, so it aims at competence and a limited form of autonomy, not relatedness. A meta-analysis of gamification in education found small positive effects on cognitive, motivational, and behavioural learning outcomes (Sailer and Homner 2020). A review of how educational games are evaluated, however, warns that their benefits are often claimed without rigorous evidence (Petri and Gresse von Wangenheim 2016). Scratch (Resnick et al. 2009) was designed to lower the entry barrier to programming by replacing typed code with blocks; we apply the same idea to prompts by representing components as selectable chips. Promptémon combines these design moves: chips for scaffolding, battles for goal-driven engagement, and a built-in judge for immediate feedback.

Table 1 compares Promptémon with K-12 AI tools and with PicPrompt. Each rating in the table is taken from the corresponding tool's own published description and evaluation; where the accessible description does not say, the cell reads "n.s.". We coded "active learning" as "yes" when the student modifies AI behaviour and "immediate feedback" as "yes" when feedback arrives within a single interaction loop. We coded "game-based" as "yes" when the experience uses goal, scoring, and progression mechanics; Promptémon uses health points, experience points, and levels. The first four tools in the table teach how a model learns from data or rewards. Cognimates has children train their own classifiers (models that sort examples into groups) inside a Scratch-like editor (Druga 2018). TrainYourSnakeAI lets middle schoolers set the rewards that train a Snake-playing agent, a program that learns from rewards (Hinojosa et al. 2025). PlayGrid lets young children watch such an agent find a path on a light-up button grid (Quiloan et al. 2023). ARtonomous has middle schoolers train virtual robot cars on an augmented reality tablet (Dietz et al. 2022). Promptémon is instead designed to teach how to instruct a model that is already trained, as students do when they use a chatbot. PicPrompt shares that focus but uses images, free-text prompts, and human judges. We are not aware of any published K-12 tool designed to teach prompt engineering through a game-based system that combines prompt assembly with immediate, AI-evaluated feedback.

# 3 Design of Promptémon

## 3.1 Learning framework

Promptémon's design rests on three complementary learning perspectives. *Constructionism* (Papert and Harel 1991; Ali et al. 2019) motivates the choice to make prompt construction (rather than reading or recall) the central activity. *Scaffolding* (Wood et al. 1976) motivates the chip system. Rather than asking students to write full prompts from a blank field, the Prompt Fusion interface decomposes a prompt into a base instruction plus three modifier categories, so students can attend to one variable at a time. Selecting from prepared options is much narrower than the open-ended building that constructionism describes, and in this version the chips never give way to free writing. *Game-based learning* (Ryan et al. 2006; Sailer and Homner 2020) motivates the battle structure, the immediate "Super Effective!" or "Not Effective…" feedback, and the use of low-stakes failure to encourage iteration. We expected these

**Table 1** Comparison of AI education tools. Ratings derived from each tool's own published description and evaluation; ages as reported by each source. n.s.: not stated in the source's accessible description.

| Tool | Learners (age) | AI concept | Active learning | Immediate feedback | No special hardware | Game-based |
|---|---|---|---|---|---|---|
| Cognimates (Druga 2018) | Children (7–14) | Machine-learning model training | Yes | Limited | Yes | No |
| TrainYourSnakeAI (Hinojosa et al. 2025) | Middle school (age n.s.) | Reinforcement learning | Yes | n.s. | Yes | n.s. |
| PlayGrid (Quiloan et al. 2023) | Young children (age n.s.) | Reinforcement learning | n.s. | n.s. | No | Yes |
| ARtonomous (Dietz et al. 2022) | Middle school (11–14) | Reinforcement learning | Yes | Limited | Yes | No |
| PicPrompt (Ngo et al. 2025; Kumar et al. 2026) | K-12 (evaluated with 22 students in grades 2–9, 4 of them in grades 6–8) | Text-to-image prompting | Yes | Yes | Yes | Yes |
| **Promptémon** | **Middle school (11–14)** | **Text prompting** | **Yes** | **Yes** | **Yes** | **Yes** |

three principles, rather than the Pokémon theme alone, to matter for learning. This study does not test them separately.

## 3.2 Game concept and learning goals

The game targets three learning goals. (1) Prompts have structured components that can be varied independently. (2) The same task may call for different prompt structures: creative tasks benefit from open framing, while constrained tasks benefit from explicit specification. (3) Prompt engineering is iterative rather than memorised. The version used in this study supports the first goal most directly. Both of its challenges are creative-writing tasks, so it does not exercise the second goal, and Section 6 discusses how its rules limit the third.

## 3.3 Gameplay mechanics

A session starts by collecting the student's first name, last name, and grade. The pre-game survey then opens inside the game window, and a guide character gives a short welcome that does not explain prompting. The student then chooses a starter from three creatures: *Stylusaur* (Style-type, depicted as an artist), *Ampwave* (Audience-type, a broadcaster), or *Masquerade* (Persona-type, a shapeshifter). The starter choice introduces the idea that there are multiple valid entry points to prompt engineering. It changes the creature's picture, type label, and attack animation, but not how attempts are judged, and it does not restrict which chips the student can use. All three chip

categories remain available regardless of starter, so a Stylusaur trainer can still use Persona and Audience chips during a battle.

Each of the two opponents presents a writing challenge: *"Write a short story about a robot pet that can only say 3 words. Make it funny!"* (Bit-Kitty) and *"Create a list of 3 cool superpowers a superhero could have. Make it exciting for kids!"* (Listle). At the start of each battle, the model writes three short base prompts (at most ten words each) that restate the challenge. It also writes three chips (at most three words each) for each of the Persona, Style, and Audience categories. The chips therefore differ from one student and one battle to the next. The student composes a prompt by selecting one base prompt from a dropdown and at most one chip per category, shown in Fig. 1; on screen, the chosen chips are listed as "Constraint Keywords". A student may also attack with no chips at all, sending the base prompt alone. (If generating the base prompts failed, the game sent the challenge text instead.) The one-chip-per-category limit prevents combinatorial overload and forces strategic choice. With three base prompts and four options in each category (no chip or one of three), a student can build 192 different prompts in a battle (Table 2).

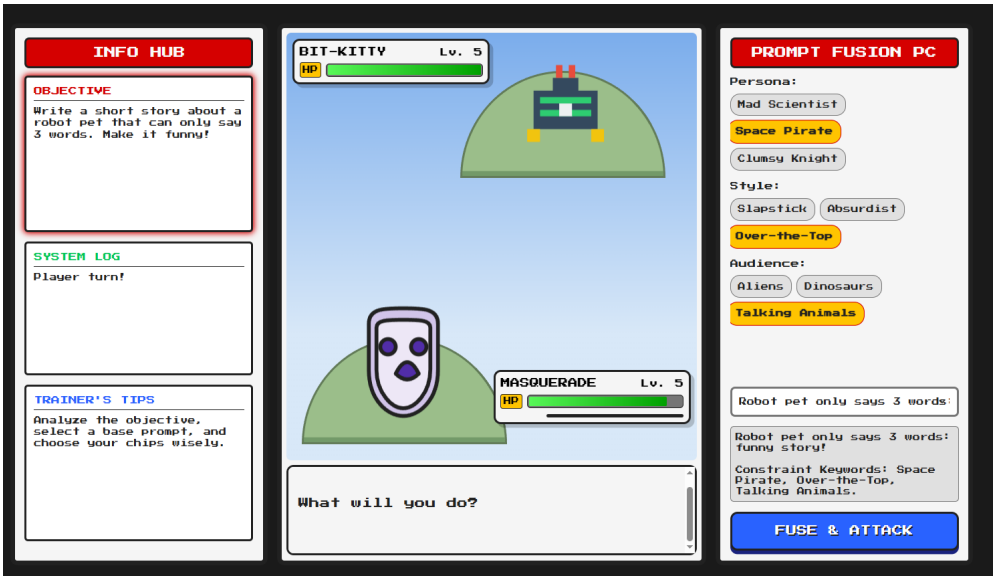



**Fig. 1** The gameplay screen during the first battle. The Info Hub (left) shows the opponent's challenge, the system log and the Trainer's Tips. The battle scene (centre) shows the opponent, Bit-Kitty, and the student's starter, Masquerade. In the Prompt Fusion panel (right) the student has picked one base prompt and one chip in each category (highlighted: Space Pirate, Over-the-Top, Talking Animals); the assembled prompt appears above the Fuse & Attack button.

Each battle is turn-based. The opponent attacks first, which lowers the player's HP (health points, 100 at the start) by a random amount of at most a quarter of the player's maximum. The student then attacks by submitting a prompt. If the judge (a second call to the model, described below) decides that the generated text does not meet the challenge, the attack has no effect, the opponent attacks again, and

the student gets another turn. If the judge decides that the text meets the challenge, the opponent is defeated at once, so a single successful attempt wins a battle. A student whose HP runs out can retry the same opponent, with newly generated base prompts and chips. Players who win earn experience points and move on to the second opponent; after both battles the post-game survey opens. Fig. 2 shows one battle as a sequence of steps.

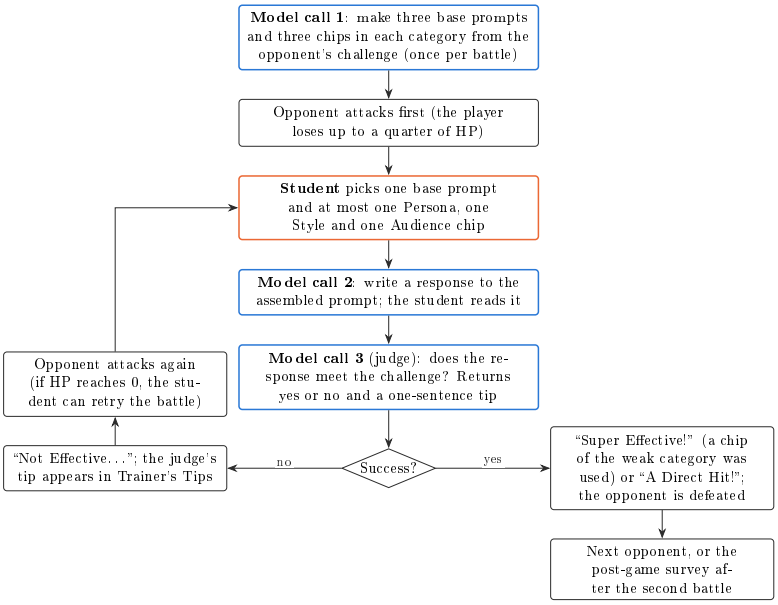



**Fig. 2** One battle in Promptémon, as implemented in the version used at the event. Blue boxes are calls to the language model; the orange box is the student's choice. The base prompts and chips are generated once per battle; a failed attempt leads to another opponent attack and another choice, and a successful attempt ends the battle.

A short example shows one turn. Facing Bit-Kitty, the student in Fig. 1 picks the base prompt "Robot pet only says 3 words: funny story!" and the chips Space Pirate (Persona), Over-the-Top (Style), and Talking Animals (Audience). The game sends the base prompt with the three chips as labelled lines, and the model's story appears in a window for the student to read. The judge then checks the story against the challenge. Because the story meets the challenge and Over-the-Top is a Style chip, the attack is "Super Effective!" (Fig. 3) and Bit-Kitty is defeated. If the story had missed the challenge, the screen would have shown "Not Effective...", the judge's tip would have appeared in the Trainer's Tips panel, and Bit-Kitty would have attacked before the student's next try.

## 3.4 Two model calls

A submitted prompt triggers two calls to the model. The *generation* call sends the student's composed prompt to Google's Gemini application programming interface

**Table 2** Parts of the game, as implemented in the version used at the event

| Element | What it is and what it does |
|---|---|
| Starters | Stylusaur (Style), Ampwave (Audience), Masquerade (Persona). Changes appearance only |
| Opponents | Bit-Kitty: robot-pet story, weak to Style. Listle: superpower list for kids, weak to Audience |
| Base prompts | Three per battle, generated by the model from the challenge; the student picks one |
| Chips | Three per category (Persona, Style, Audience), generated by the model for each battle; at most one per category |
| Prompt sent | The base prompt followed by the chosen Audience, Style and Persona lines |
| Judge | A second call to the same model; returns success or failure and a one-sentence tip |
| Outcome | Failure: "Not Effective…". Success: "A Direct Hit!", or "Super Effective!" if a chip from the opponent's weak category was used; either defeats the opponent |
| Opponent attack | Before the first attempt and after each failure; at most a quarter of the player's maximum HP |

(API), the service through which the game exchanges text with the model. The game then displays the response so the student can read it. A second *evaluation* call sends the response and the opponent's challenge text to the same model in a separate request. It tells the model to act as "a fair but encouraging game judge for a middle school audience" and to judge only the generated text, not the student's prompt. It also tells the judge to focus on "the spirit of the challenge, not minor mistakes" and to return whether the output met the main goal, together with a one-sentence tip. Splitting generation from evaluation lets the student see real model output and receive a separate verdict on it. The judge returns a yes-or-no decision with no numeric score. That decision is game feedback and has not been validated as a measure of a student's prompting.

The same model generates and judges. A model asked to judge text may favour text like its own; in summarisation tasks, LLM judges recognised and preferred their own outputs (Panickssery et al. 2024). We did not check the judge's decisions against human raters, and both the feedback a student received and the moment a battle ended depended on those decisions. One more call per battle generates the base prompts and chips. The full instructions for all three calls are given in Supplementary Information S1.

## 3.5 Feedback system

Feedback works on two layers (Fig. 3). A *visceral* layer uses sound and overlay text ("Super Effective!", "A Direct Hit!" or "Not Effective…") so success or failure is felt immediately. An *instructional* layer in the Trainer's Tips panel shows the judge's one-sentence tip, which explains why the output met or missed the challenge and how to improve it. When an attack is "Super Effective!", a fixed message replaces the tip and names the weakness the student used. Any success already ends the battle, so this message praises the choice after the fact and does not change the outcome. After a failed attempt the judge writes the tip, so it may name a specific change to make. That can steer a student towards a known answer instead of open experimentation. This is a trade-off between guiding the student and preserving experimentation, which we discuss in Section 6.3.

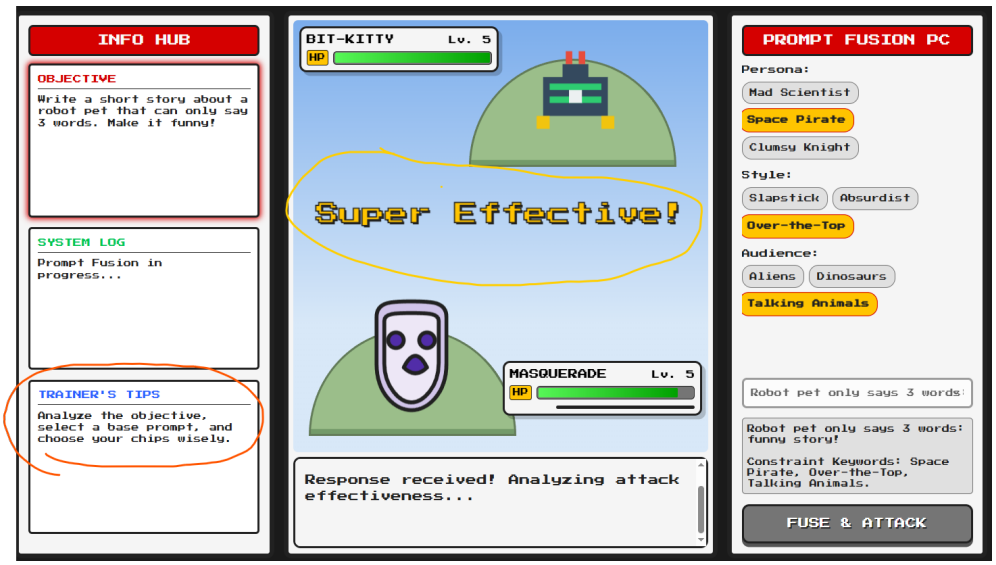



**Fig. 3** Two-layer feedback. The "Super Effective!" overlay (circled, centre) gives immediate feedback on a successful attack that used a chip from the opponent's weak category. The Trainer's Tips panel (circled, left) is where the judge's one-sentence tip appears after an attempt; in this screenshot it still shows the default instruction.

### 3.6 Implementation

Promptémon is a single-page web application written in HTML5, JavaScript, and CSS (the standard languages of web pages), with no installation or special hardware required. The version used in this study called Google's Gemini API (model `gemini-2.0-flash-exp`) for all three kinds of call, but all calls pass through one function, so swapping in another LLM is straightforward. Health bars, colour-coded chip categories, and character animations follow conventions students already recognise from popular battle games. The two challenges are plain text in the file, so a teacher can set new writing tasks by editing them; the base prompts and chips are then generated from the new text. The game also wrote a record of each attempt to an online database, which is no longer available (Section 4).

## 4 Methods

### 4.1 Approach

This work is an exploratory, single-session design study. Our main goal is to describe the formative use of Promptémon and to reflect on what worked, not to claim generalisable causal effects. We describe pre-to-post changes in questionnaire responses, using both rating scales and the students' own words, rather than in-game battle results. TrainYourSnakeAI, a middle-school AI tool, was evaluated in a similar way, by surveying students' AI knowledge before and after they used it (Hinojosa et al. 2025). We used a within-subject design (each student answered before and after play) with

two data sources: (1) a pre-game questionnaire and (2) a post-game questionnaire that repeated some items. The game also logged every prompt submission, but the cloud project that hosted the database is no longer active and its records could not be recovered, so this article reports the questionnaires only. We chose within-subject pre/post comparison over comparing two separate groups because each participant only had a single window of seven to ten minutes with the tool.

The short visit shaped the instruments as well as the findings. With seven to ten minutes per student, including both surveys, there was no time for a performance task, such as writing a prompt for a new problem, or for an interview. We therefore kept the questionnaires short, with rating items, one multiple-choice item, and a few open questions. As a result, the study can describe how students rated their own confidence and understanding, but not how well they prompt.

## 4.2 Participants and setting

The study took place on 14 and 15 August 2025 at the AI Literacy Expo, a two-day event with two six-hour sessions. It was hosted at a public middle school in a small Texas city. The school follows a curriculum focused on science, technology, engineering, and mathematics (STEM). It is a regular public school whose schedule emphasises STEM electives, not a selective magnet programme (a specialist US school that chooses its students). Students from the school's after-school programme participated, and we received 60 forms before play and 60 after. The pre-game form recorded 24 sixth graders (40%), 15 seventh graders (25%), and 21 eighth graders (35%). Students in these grades are usually 11 to 14 years old, but the forms did not ask for age. The post-game form recorded 26, 16, and 18; every matched student gave the same grade on both forms, so the difference lies among the unmatched forms. Five tables, each demonstrating a different AI application, were arranged around the room. Students worked in groups of five or six and rotated between tables every seven to ten minutes. We did not record the order in which groups visited the tables, so many students had already used other AI activities before they reached Promptémon. The study was approved by the Institutional Review Board of The University of Texas Permian Basin (protocol 2025-022), and we obtained both parental consent and student assent (see Declarations).

## 4.3 Procedure

Each student spent about seven to ten minutes with Promptémon. A team member gave a short spoken introduction to the station; its content was not recorded. By the team's estimate, the pre-game questionnaire and the in-game welcome together took about two minutes, the gameplay portion took five to seven minutes (up to two battles), and the post-game questionnaire took about one minute. The time stamps on the forms give a partial check. Among the 53 matched students, the median time between submitting the two forms was 8.7 minutes (middle half 7.0 to 11.4 minutes). Eighteen of the 53 gaps were over ten minutes, the longest being 14.3 minutes. Three were under five minutes, and in one of those the post-game form was stamped before the pre-game form, so that match may be wrong. During the first battle, after the

**Table 3** Survey items, as worded on the forms. Rating items used a 1–5 scale labelled only at the ends, except the revision item, which had no labels

| Form | Item | Response format |
|---|---|---|
| Pre | How confident are you in your ability to get a useful or creative response from an AI like ChatGPT or Google Gemini? | 1 = Not confident at all ... 5 = Very confident |
| Pre | When you use an AI, how often do you change your prompt or try again if you don't get the result you want on the first try? | 1–5, no labels |
| Pre | I believe that changing the persona, style, or intended audience of a prompt can significantly change the AI's answer. | 1 = Strongly disagree ... 5 = Strongly agree |
| Pre and post | To get the BEST answer from an AI, what is the best way to write your prompt? | One of four options (Table 4) |
| Pre | In your own words, what does it mean to give an AI a 'good' or 'effective' prompt? | Open text |
| Post | After playing Promptémon, how confident are you now in your ability to get a useful or creative response from an AI? | 1 = Not confident at all ... 5 = Very confident |
| Post | The game made me want to try different combinations of 'chips' to see what would happen. | 1 = Strongly disagree ... 5 = Strongly agree |
| Post | The feedback from the game (like 'Super Effective!' or the Trainer's Tips) helped me understand how to make my prompts better. | 1 = Strongly disagree ... 5 = Strongly agree |
| Post | I now have a better understanding of how using a specific persona, style, or audience can change an AI's response. | 1 = Strongly disagree ... 5 = Strongly agree |
| Post | How did you find the experience of learning through a game compared to a traditional classroom lesson? Please explain your answer. | Open text |
| Post | What was your favorite part of the 'Prompt Fusion' system (combining a base prompt with chips)? Why? | Open text |
| Post | If you were to explain to a friend how to write a good prompt for an AI, what would you tell them now that you have played Promptémon? | Open text |

opponent's opening attack, a five-step on-screen tour pointed out the objective, the base prompt menu, the chips, the assembled prompt, and the attack button. Beyond the spoken introduction and the in-game tips, students played without coaching, so that we could observe how the game itself communicates the mechanics.

Both questionnaires were online forms shown inside the game window. Table 3 lists every item with its wording and response format. The items were written for this study and were not taken from a validated scale. The pre-game questionnaire used five-point rating items measuring (a) confidence in getting useful AI responses, (b) how often students typically revise prompts when using AI tools in general, and (c) belief that persona, style, and audience can change AI output. Item (b) was asked before play only. The questionnaire also included one multiple-choice item on the "best way to write your prompt" (with four candidate strategies). Its one open-ended item asked: "In your own words, what does it mean to give an AI a 'good' or 'effective' prompt?" The post-game questionnaire repeated the multiple-choice item and the confidence item (reworded to refer to the game), so that we could compute pre/post change. It replaced the belief item with an item on understanding, and added two agreement items on the game and three open-ended items. Students also gave their grade and name on each form so that the two forms could be matched.

## 4.4 Data analysis

We received 60 forms before play and 60 after. We matched pre and post forms by first and last name and then removed the names. Fifty-three students could be matched. Six pre-game and five post-game forms had no partner, and one name that appeared once on the pre-game form and twice on the post-game form was excluded. Most unmatched forms had no similar name on the other form, which suggests that some students completed only one of the two forms; we did not try matching on similar spellings. Every matched pair reported the same grade on both forms. For rating and multiple-choice items we report distributions for all 60 respondents and, where an item was asked in the same way both times, the within-student change for matched students.

Confidence was the only rating item asked in comparable form before and after play. We tested the change in matched students' ratings with a two-sided exact sign test, which compares the number of students whose rating rose with the number whose rating fell. It does not assume that rises and falls are spread evenly or that the steps between scale points are equal, which suits a five-point rating. It drops the 19 unchanged pairs and uses the 34 students whose rating changed. As a secondary check we report the two-sided Wilcoxon signed-rank test, which also uses those 34 pairs; it ranks the size of each change and so treats the steps between ratings as equal. We also report the standardised mean change $d_z$ (the mean change divided by the standard deviation, SD, of the changes) with a bootstrap 95% confidence interval (CI), estimated by resampling the matched students 10,000 times. For the multiple-choice item we coded the option "Tell the AI its job ..., what you want it to do ..., and who the answer is for ..." as the structured strategy. We tested the matched change with a two-sided exact McNemar test, which suits a paired yes-or-no outcome, using the 52 matched students who answered the item both times. The sign test and the McNemar test are the two primary tests. We applied the Holm step-down adjustment, which corrects the $p$ values for running two tests, with $\alpha = 0.05$. We chose the tests after collecting the data and did not pre-register the study, so all inference is exploratory. The belief item and the understanding item were worded differently, so we report them separately and do not compute a change.

For the open-ended items we used two simple, repeatable methods. First, following the counting form of content analysis (Hsieh and Shannon 2005), a fixed word list counted answers that named a prompt component (persona, style, audience, tone, "who it's for" and its variants, or "act like/as"). A second list counted answers that advised being specific (words containing "specif", "detail" or "descri", words beginning "clear", or the word "exactly"). The word lists are given in Supplementary Information S2, and the script is in Supplementary Code 1. Second, we sorted the answers to the game-versus-classroom item into three groups (favours the game; favours a classroom lesson; mixed, unclear, or not a comparison) using a written rule fixed before sorting. We made two sorting passes with an AI assistant (Claude Opus 5.5, Anthropic). The first was made by the working session that was helping us with the analysis, so it had already seen the study's other results. The second was a separate session given only the rule and the answers, without names or grades, and it could not see the first pass. They agreed on 59 of 60 answers (Cohen's $\kappa = 0.97$, a measure

of agreement that allows for chance), and the rule placed the one disagreement in the mixed group. Because both passes used the same assistant, the agreement shows only that the rule was applied consistently. No human coder sorted the answers independently, so we do not know how people would sort them. The sorted answers are in Supplementary Data 1, so readers can check the sorting. We also used AI assistants to check numbers and review drafts (see Declarations). We quote students by grade only.

# 5 Results

Before play, students also reported how often they change a prompt or try again when an AI's first answer is not what they want. On this unlabelled 1–5 scale, the counts for ratings 1 to 5 were 4, 9, 23, 18, and 6. Twenty-four of 60 chose 4 or 5, and 47 of 60 chose 3 or higher. With no labels on the form, we do not know how students read this scale. The item was not repeated after play.

## 5.1 RQ1: confidence and self-reported understanding

***Confidence.***

On the 1–5 scale, the number of students rating their confidence 4 or 5 rose from 29 of 60 (48.3%) before play to 47 of 60 (78.3%) after play. The counts for ratings 1 to 5 were 0, 2, 29, 22, and 7 before play and 0, 3, 10, 21, and 26 after. The top rating went from 7 to 26 students. Among the 53 matched students, 26 rated their confidence higher after play, 8 lower, and 19 the same (Fig. 5). The two-sided exact sign test indicated a statistically significant increase ($n = 53$, $p = .0029$, Holm-adjusted $p = .0059$); the two-sided Wilcoxon signed-rank test agreed ($W = 100$, $p = .0004$, 34 non-zero pairs). Leaving out the three pairs with under five minutes between forms gave the same result (24 up, 8 down, 18 same; sign test, $n = 50$, $p = .0070$). Treating the steps between ratings as equal, mean confidence rose from 3.58 to 4.13 (mean change 0.55, SD 1.01; $d_z = 0.54$, 95% CI 0.30 to 0.83). This is a figure from one group measured twice, so it cannot be compared with effect sizes from studies that have a comparison group. We note one important caveat: confidence is a self-report and could reflect novelty effects rather than durable skill (Clark 1983; Tsay et al. 2020). The post-game item also named the game. Because many students had visited other AI stations first, all or part of this rise may come from the event as a whole rather than from Promptémon. Seven of the 53 matched students started at the top of the scale and could not rise. Full distributions are in Fig. 4.

***Belief that prompt elements affect output.***

Before play, 34 of 60 students (56.7%) agreed (4 or 5) with the item "I believe that changing the persona, style, or intended audience of a prompt can significantly change the AI's answer". Their ratings from 1 to 5 were 0, 5, 21, 23, and 11. After play, 44 of 60 (73.3%) agreed that they now had a better understanding of how persona, style, or audience can change an AI's response (ratings 0, 2, 14, 26, and 18). Because the two items ask different things (a belief before play, a gain in understanding after), we do not treat the difference as a change. The post-game figure shows that most students

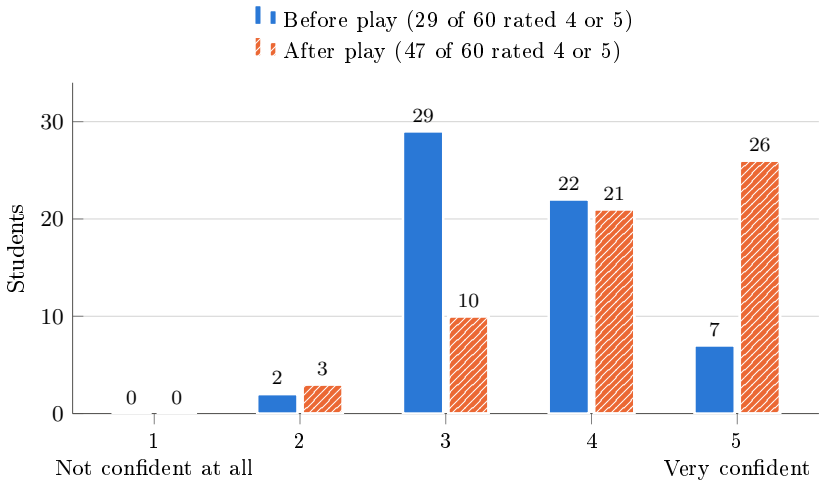



**Fig. 4** Confidence in getting useful AI responses, before and after play (all 60 respondents on each form). Ratings of 4 or 5 rose from 29 of 60 to 47 of 60.

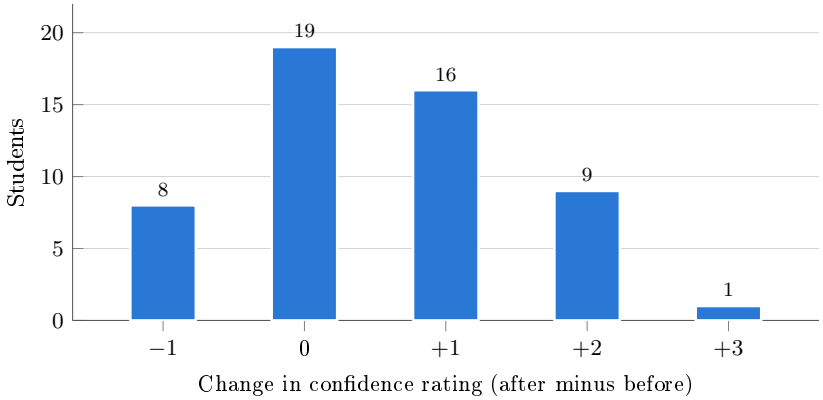



**Fig. 5** Change in each matched student's confidence rating (after minus before; $n = 53$). Twenty-six students' ratings rose, 8 fell, and 19 did not change.

said they understood these three parts better, although the item did not name the game. It does not show what they would do differently.

***Strategy choice.***

The multiple-choice item asked which of four strategies produces the best AI answer. Table 4 gives the four options in the forms' wording and order. The structured option ("Tell the AI its job . . . ") is the most explicit and the longest, and it is the closest to the game's own categories. The "let it decide" option is the most open-ended.

Before playing, 34 of 60 forms chose the structured option (one form left the item blank); after playing, 29 of 60 did, while the open-ended option rose from 14 of 60

**Table 4** Choice of the best way to write a prompt, in the forms' order (all respondents, $n = 60$ per form)

| Option (verbatim) | Pre | Post |
|---|---|---|
| Ask a really short question, like "tell me about dogs." | 8 | 8 |
| Use big, smart-sounding words to show the AI you're serious. | 3 | 4 |
| Give the AI a topic, like "write about space," but let it decide everything else. | 14 | 19 |
| Tell the AI its job (like "act as a scientist"), what you want it to do (like "make a list"), and who the answer is for (like "a 5th grader"). | 34 | 29 |
| No answer | 1 | 0 |

to 19 of 60. Among the 52 matched students who answered both times, 31 chose the structured option before and 27 after. Twenty-six chose it both times, 5 moved away from it, and 1 moved to it (two-sided exact McNemar test, $p = .22$, Holm-adjusted $p = .22$). Among the same 52 students, the open-ended option went from 12 to 18 (not tested). We interpret this shift cautiously. Two readings are plausible. Some students may have learned that the right amount of detail depends on the task. One sixth grader advised a friend to use "a descripted response but not to much and let the AI the rest". It is equally possible that the battles taught some students that less structure works, because a few short chips produced a whole story. Another sixth grader wrote that you can "just give the AI a couple topics and it will give them a whole entire story". That second reading is the opposite of the lesson the design intends. Because the change is not statistically significant ($p = .22$) and rests on only six students who moved into or out of the structured option, we cannot distinguish these readings from the data we have.

***Students' own words.***

The open answers show that the game's terms did not often carry into students' own words. The pre-game and post-game open items asked different questions, and the answers should be read with that limit in mind. Before play, students were asked what it means to give an AI a good prompt. Many answers gave general advice, such as "A good prompt is a detailed prompt" (sixth grader). Sixteen of 60 advised being specific, detailed, descriptive, exact, or clear, and only one named a component ("like who its for and what its about", seventh grader). After play, students were asked what they would tell a friend about writing a good prompt. Fourteen of 60 advised being specific, detailed, descriptive, exact, or clear, and 5 of 60 named a component. One eighth grader wrote: "I would tell them to be specific about specific things like, the tone, the audience, and what you want exactly when making a prompt." A second eighth grader wrote: "I would tell them to make sure to tell the AI who the target audience is and to think about who will read it and how they need to accomodate their writing." One of the five ("act like a scientists", eighth grader) echoes the example in the multiple-choice item rather than a game label. Several other post-game answers were about the game rather than about prompting ("i would tell them it was fun", sixth grader). Because the game labels its three chip categories *Persona*, *Style*, and *Audience*, answers that use these words may partly repeat the labels rather than show understanding. We address this directly in Section 6.3.

## 5.2 RQ2: experimentation and iterative refinement

***Self-reported motivation to experiment.***

Asked whether the game made them want to try different chip combinations, 28 of 60 students (46.7%) agreed or strongly agreed, as shown by the first set of bars in Fig. 6. The counts for ratings 1 to 5 were 4, 6, 22, 19, and 9. For a session of seven to ten minutes, this is a modest share. Ten students chose 1 or 2, and 22 (36.7%) chose the middle of the scale. We cannot tell whether this reflects the short session or low interest.

***What students said about the chips.***

Without the turn logs, we cannot report how often students changed chips between attempts. The favourite-part answers give a partial view. Some students named experimentation ("combining it allowed me to experiment", eighth grader). Others named the effect of a component ("My favorite part was picking who the audience was because it showed how AI can change what it says depending on who its for", seventh grader). Three answers named the chip categories; a fourth matched the word list only because it praised "The style of the game". Others found the interface hard: one sixth grader wrote, "Honestly, I just pressed the attack button because it looked super complex."

***Feedback helpfulness.***

Forty-four of 60 students (73.3%) agreed or strongly agreed that the feedback helped them understand how to improve their prompts, as shown by the second set of bars in Fig. 6. The counts for ratings 1 to 5 were 0, 3, 13, 30, and 14. The item names both the "Super Effective!" overlay and the Trainer's Tips, so it cannot tell us which layer students valued. Agreement also does not show that students used the feedback.

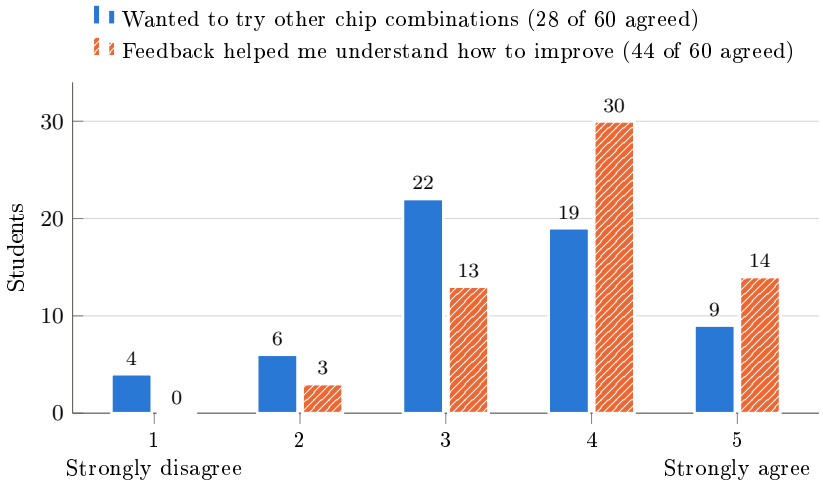



**Fig. 6** Post-game agreement with the two RQ2 items ($n = 60$ each). Counts of 4 or 5 (agree): 28 of 60 for wanting to try other chip combinations and 44 of 60 for the feedback helping.

**Table 5** Summary of survey results. Unpaired counts use all 60 forms in each set, and one pre-game strategy answer was blank; tests use matched students. Both tests are two-sided and exact, with Holm-adjusted $p$ in brackets

| Item | Before | After | Matched change | Test, $p$ |
|---|---|---|---|---|
| Confidence (Before and After count ratings of 4 or 5; matched change is on the full 1–5 scale) | 29 / 60 | 47 / 60 | 26 up, 8 down, 19 same ($n = 53$) | Sign test, .0029 [.0059] |
| Chose the structured strategy | 34 / 60 | 29 / 60 | 31 → 27 ($n = 52$) | McNemar, .22 [.22] |
| Chose "let it decide everything else" | 14 / 60 | 19 / 60 | 12 → 18 ($n = 52$) | Not tested |
| Persona, style, audience: belief (pre) / understanding (post), agreed | 34 / 60 | 44 / 60 | Different items, not compared | – |
| Feedback helped me understand how to improve, agreed | – | 44 / 60 | Post only | – |
| Wanted to try other chip combinations, agreed | – | 28 / 60 | Post only | – |
| Open answer named a prompt component | 1 / 60 | 5 / 60 | Different questions, not compared | – |
| Favourite part named the chip categories | – | 3 / 60 | Post only | – |
| Game or classroom lesson (open answer) | – | 30 game, 14 lesson, 16 mixed | Post only | – |

***Learning through a game compared with a classroom lesson.***

Answers to this item ran both ways. Thirty of 60 students favoured the game, 14 favoured a classroom lesson, and 16 gave mixed, unclear, or off-topic answers. Students who favoured the game mentioned interaction, engagement, and feedback. One eighth grader wrote, "I liked learning through a game better than a traditional classroom lesson because it was much more interactive"; another wrote, "I found that the game helped me be more engaged and get specific personal feedback on what i did". Some students who favoured a lesson valued a person's explanation. An eighth grader wrote, "I find it better to learn through a traditional classroom because it is better when someone is explaning it to you"; a sixth grader wrote, "I think a traditional classroom would be better, because understanding a human is easier than understanding an AI". One sixth grader wrote, "it was fun but it didnt teach me anything".

## 5.3 Pre and post survey summary

Table 5 summarises the comparisons. Confidence rose and the change was statistically significant (two-sided exact sign test, $n = 53$, 34 non-zero pairs, $p = .0029$, Holm-adjusted $p = .0059$). The choice of the structured strategy did not rise, and the change was not statistically significant (two-sided exact McNemar test, $n = 52$, $p = .22$), so we cannot rule out that it occurred by chance. This result is consistent with the discussion in Section 5.1: after playing, some students moved towards the open-ended option rather than the structured one. From one short session, with only six students changing either way, the data show no evidence of a shift towards the structured strategy.

# 6 Discussion

## 6.1 What the findings say about RQ1

For RQ1, the data support a narrow claim. A single short session was followed by a rise in students' self-reported confidence: 26 of the 53 matched students rated themselves higher, and 8 rated themselves lower. Most students also agreed afterwards that they understood better how persona, style, and audience change a response. We make no claim of prompting skill or learning, and beyond that self-report the survey gives no evidence that understanding improved. On the one multiple-choice item asked the same way twice, choice of the structured strategy did not rise, and in the open answers only 5 of 60 post-game answers named a component. We therefore have no evidence that students would do anything differently, even though their confidence rose. This reading is consistent with work showing that how much people feel they have learned can differ from what they have learned (Sitzmann et al. 2010). It is also consistent with the possibility that the novelty of a game raised confidence on its own (Tsay et al. 2020). The game put three parts of a prompt on screen and named them. Whether students learned those parts in the session, or could use them on new tasks, remains untested.

A design feature may also explain the strategy result. In Promptémon, a few playful chips such as "Space Pirate" produce a complete story, and the judge checks the story against a light-hearted challenge. A student could reasonably conclude that short inputs are enough and that the model will fill in the rest, which is close to the "let it decide everything else" option. The structured option on the survey describes purposeful choices (a job, a task, a reader), while the game's chips are written to be playful. A version of the game that asks students to match a chip to a stated purpose, such as a specific reader, would test whether that gap matters. The three categories also suit the game's creative-writing challenges better than everyday school tasks. When students ask an AI to explain a topic or check their work, constraints and accurate subject terms may matter more than persona or style, and the game does not include them.

## 6.2 What the findings say about RQ2

For RQ2 (experimentation), fewer than half of the students (28 of 60) said the game made them want to try other chip combinations, and without the logs we cannot see what they actually did. The design itself may explain part of this. Because one successful attempt ends a battle, a student whose first prompt works has no reason inside the game to vary it; further attempts happen only after a failure. The "Super Effective!" message does not change this, because it appears only on a success that has already ended the battle. The game therefore allowed experimenting but did not require it. Forty-four of 60 students agreed that the feedback helped them understand how to improve, and one valued its "specific personal feedback". These self-reports do not show that students used the tips or wrote better prompts. Half of the students (30 of 60) favoured the game over a classroom lesson. Fourteen favoured a lesson, some because they valued a person's explanation, and 16 gave mixed or unclear answers.

## 6.3 Limitations

**Single site, no control:** one Texas STEM-focused middle school is not generalisable, and we cannot compare Promptémon to a non-game alternative. Students visited other AI stations in an order that was not recorded, so their answers may reflect the wider event rather than this game. Participants came from an after-school programme at an AI event, and the forms recorded nothing about them beyond grade, such as prior AI use. Students moved in groups of five or six, so they may have talked to or watched one another, although our tests treat each student as independent. **Self-report dominance:** every rating and multiple-choice item is a self-report, and the post-game confidence item named the game; novelty effects cannot be ruled out (Clark 1983; Tsay et al. 2020). **Same model for generation and evaluation:** the design uses Gemini for both the response and the judgement. In summarisation tasks, LLM judges have favoured their own outputs (Panickssery et al. 2024), so the same bias may affect the judge, and we did not check the judge's decisions against human raters. A follow-up study should add human or alternative-model judging on a sample of responses. **No measure of learning or transfer:** we checked only whether survey answers chose or repeated the persona/style/audience vocabulary, not whether students understood it or used it during or after play. The structured survey option uses the same ideas as the game's labels, so choosing it partly tests recognition. A stronger evaluation would ask students to write a free-form prompt for a new task with no chip interface. **Hint bias:** after a failed attempt, the judge's tip can point students to a particular change, which may help them but risks confounding the experimentation measure. In the other direction, the judge saw only the generated text, not the student's prompt or chips, so its tips could not comment directly on those choices. Because one success ended a battle, and "Super Effective!" replaced the tip with a fixed message, a student who won had no turn in which to act on the feedback. **Single short session:** we cannot speak to retention (whether any change lasts). A team member also gave an unrecorded spoken introduction before play, which may have influenced responses. The times given for each part of the session were team estimates; only the gap between the two forms was measured. **No log analysis:** the database of attempts is no longer available, so we cannot report how students changed their prompts, and we did not assess the quality of the prompts they built. **Unrecorded options:** the base prompts and chips were generated anew for each battle, so students saw different options, and without the logs we do not know which ones. **Instrument:** the questionnaire was short and not validated. The understanding item after play was worded differently from the belief item before play, and the open questions differed between the two forms. Rating items labelled only their end points, and the pre-game revision item had no labels. The post-game agreement items were all worded positively, and the confidence item named the game, which can encourage favourable ratings. Several items combined two or three ideas in one question. The structured multiple-choice option was the longest of the four, and it supplied the wording that one post-game answer echoed. **Analysis and missing students:** matching by exact first and last name left 7 of the 60 pre-game forms and 7 of the 60 post-game forms unpaired. Because the post-game form opened only after

both opponents were beaten, students who ran out of time are missing from the post-game and matched data, and we do not know how many there were. Unpaired totals ($n = 60$) include forms that the paired tests ($n = 53$ and $n = 52$) leave out. The two tests were chosen after data collection and are exploratory. Only six of the 52 matched students changed into or out of the structured option, so the McNemar test could easily miss a real change. Seven matched students started at the top rating and could not rise, and $d_z$ and the Wilcoxon test treat the steps between ratings as equal. The word counts are a coarse measure of the open answers: they miss other ways of saying the same thing, and they can match the wrong thing, as with "The style of the game". No human coder sorted the classroom item independently. Both sorting passes came from the same AI assistant, and the first had seen the study's other results, so $\kappa = 0.97$ shows consistency, not agreement between people. The rule also counted general praise for learning through the game as favouring the game, even when the student did not compare it with a lesson, so the count of 30 is generous.

# 7 Conclusion and future work

In this exploratory study at one middle school event, students played Promptémon and answered short surveys during a station visit of about seven to ten minutes. Their self-reported confidence in getting useful AI responses rose afterwards, but with no comparison group, and with other AI stations at the same event, we cannot attribute that rise to the game. Choice of the structured prompting strategy did not rise, and few students used the component terms in their own words. We see three directions for future work. First, we plan to test whether playing Promptémon helps students use persona, style, and audience in free-form prompts on new tasks where the chip interface is not available. Second, we plan repeated sessions with a delayed post-test to assess retention, a comparison with a classroom lesson on the same content, and larger samples from more than one school. Third, we plan three design changes suggested by these results. The game would keep a battle going after a success so that students compare prompts, ask students to justify a chip choice, and use a different model, or people, to judge. The source code and the prompts that drive the game are shared with this article so that researchers and teachers can inspect, test, and adapt the tool.

**Supplementary information.** Supplementary Information S1 gives the full instructions the game sends to the model (generating the base prompts and chips, generating responses, and judging). Supplementary Information S2 gives the word lists and the written rule used for the open-ended answers. Supplementary Information S3 reproduces the two survey forms. Supplementary Data 1 is an Excel workbook of the raw and processed data. Supplementary Code 1 is a zip file of the data as CSV files, the analysis scripts and the game's source code.

**Acknowledgements.** The authors thank the participating students, their families, and the school staff who hosted the AI Literacy Expo.

## Declarations

**Funding.** This work was supported in part by the U.S. Department of Education through the Hispanic-Serving Institutions STEM (HSI-STEM) program at The University of Texas Permian Basin (grant P031C210156).

**Competing interests.** The authors declare no competing interests.

**Ethics approval.** The study was approved by the Institutional Review Board of The University of Texas Permian Basin under expedited review (protocol 2025-022, approved 30 July 2025), as part of the project "Developing AI Driven Learning Tools to Promote AI Literacy". The data were collected on 14 and 15 August 2025, within the approval period. All study procedures followed the university's policies for research involving human participants (Handbook of Operating Procedures, Part 2, Section 4). These policies are guided by the Belmont Report and apply the U.S. rules for protecting research participants (45 CFR Part 46).

**Consent to participate.** Informed consent was obtained from the parents or legal guardians of all participants, and all participants gave their assent.

**Consent for publication.** The article and its supplementary files contain only aggregate results, short anonymous quotations labelled by grade, and survey responses with names and submission times removed. No names, images or other information that could identify a participant are published. The screenshots show the game interface only.

**Data availability.** The survey data, instruments and code behind this study are shared as supplementary files with this article. Supplementary Data 1 (Excel workbook) holds the raw survey data, with names and submission times removed: the ratings and multiple-choice answers from the pre-game and post-game forms (60 rows each) and the 240 written answers. It also holds the processed data: the 53 matched pairs, the sorted classroom answers, the word-list matches, and the counts and test results reported here. Supplementary Code 1 (zip file) holds the same data as CSV files, the scripts that reproduce the reported statistics from those files, and the game's source code. The survey instruments are in Supplementary Information S3, the instructions the game sends to the model are in S1, and the word lists, the sorting rule and the sorting instruction are in S2. Written answers are stored apart from the ratings, with separate IDs, so that a recognisable answer cannot be linked to the same student's other responses. Students' names were used only to match forms and are not shared, so the matching itself, the breakdown of the unmatched forms and the check of similar names cannot be rerun from the shared files. The game's database of attempts is no longer available, so no gameplay logs can be shared.

**Code availability.** The game is a single HTML file. Its source code is in Supplementary Code 1, with the access keys used at the event and the addresses of the live survey forms and database removed, so it needs a Gemini API key to run. The analysis scripts are in the same file. The full instructions the game sends to the model are given in Supplementary Information S1.

**Author contributions.** P.L. designed and built the game, collected and analysed the data, and wrote the first draft of the manuscript. P.K. conceived and supervised the study, secured funding, and revised the manuscript. K.S. contributed equally to analysing the data and revising the manuscript. P.L. and P.K. read and approved the final manuscript.

**Use of AI tools.** An AI assistant was used to sort the classroom answers (see Methods). The authors also used AI assistants (Claude, Anthropic; Gemini, Google) to check reported numbers against the data, screen the written answers for names before sharing them, suggest revisions, and review drafts. The authors reviewed all text and take full responsibility for the content.

## References


AI4K12 (2020) AI4K12: Five Big Ideas in artificial intelligence. URL https://ai4k12.org, accessed 24 October 2026

Ali S, Payne BH, Williams R, et al (2019) Constructionism, ethics, and creativity: developing primary and middle school artificial intelligence education. In: International Workshop on Education in Artificial Intelligence K-12 (EDUAI '19), URL https://www.media.mit.edu/publications/constructionism-ethics-and-creativity/

Ali S, Ravi P, Moore K, et al (2024) A picture is worth a thousand words: co-designing text-to-image generation learning materials for K-12 with educators. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 23260–23267, https://doi.org/10.1609/aaai.v38i21.30373

Arrington C, McVey D, Mativo J, et al (2025) Enhancing teacher effectiveness with AI based prompt engineering: a proof of concept. Journal of STEM Education: Innovations and Research 26(2):5–10. https://doi.org/10.63504/jstem.v26i2.2711

Chen E, Wang D, Xu L, et al (2024) A systematic review on prompt engineering in large language models for K-12 STEM education. arXiv preprint arXiv:2410.11123, https://doi.org/10.48550/arXiv.2410.11123

Clark RE (1983) Reconsidering research on learning from media. Review of Educational Research 53(4):445–459. https://doi.org/10.3102/00346543053004445

Dietz G, King Chen J, Beason J, et al (2022) ARtonomous: introducing middle school students to reinforcement learning through virtual robotics. In: Proceedings of the 21st Annual ACM Interaction Design and Children Conference. ACM, pp 430–441, https://doi.org/10.1145/3501712.3529736

Druga S (2018) Growing up with AI. Cognimates: from coding to teaching machines. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, URL https://hdl.handle.net/1721.1/120691

Hinojosa C, Kumar P, Rajarajan PD, et al (2025) TrainYourSnakeAI: a novel tool to teach reinforcement learning to middle school students. In: Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1. ACM, pp 506–512, https://doi.org/10.1145/3641554.3701907

Hsieh HF, Shannon SE (2005) Three approaches to qualitative content analysis. Qualitative Health Research 15(9):1277–1288. https://doi.org/10.1177/1049732305276687

Kumar P, Lim P, Sotelo A, et al (2026) PicPrompt: a game-based tool for teaching prompt engineering through text-to-image generation in K-12 settings. ASEE Computers in Education Journal 15(4):108–130. https://doi.org/10.18260/B4B8-9A-60472

Laupichler MC, Aster A, Schirch J, et al (2022) Artificial intelligence literacy in higher and adult education: a scoping literature review. Computers and Education: Artificial Intelligence 3:100101. https://doi.org/10.1016/j.caeai.2022.100101

Lee I, Ali S, Zhang H, et al (2021) Developing middle school students' AI literacy. In: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education. ACM, pp 191–197, https://doi.org/10.1145/3408877.3432513

Li MZ (2024) Using prompt engineering to enhance STEM education. In: 2024 IEEE Integrated STEM Education Conference (ISEC). IEEE, https://doi.org/10.1109/ISEC61299.2024.10664775

Lo LS (2023) The CLEAR path: a framework for enhancing information literacy through prompt engineering. The Journal of Academic Librarianship 49(4):102720. https://doi.org/10.1016/j.acalib.2023.102720

Long D, Magerko B (2020) What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, pp 1–16, https://doi.org/10.1145/3313831.3376727

Ma Q, Jain A, Kim J, et al (2025) ImaginAItion: promoting generative AI literacy through game-based learning. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. ACM, pp 1–9, https://doi.org/10.1145/3706599.3719844

Ng DTK, Su J, Leung JKL, et al (2024) Artificial intelligence (AI) literacy education in secondary schools: a review. Interactive Learning Environments 32(10):6204–6224. https://doi.org/10.1080/10494820.2023.2255228

Ngo J, Lim P, Kopp G, et al (2025) WIP: PicPrompt – a novel social prompt engineering game based tool for informal learning in K-12. In: 2025 IEEE Frontiers in Education Conference (FIE). IEEE, https://doi.org/10.1109/FIE63693.2025.11328282

Panickssery A, Bowman SR, Feng S (2024) LLM evaluators recognize and favor their own generations. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), pp 68772–68802, https://doi.org/10.52202/079017-2197

Papert S, Harel I (1991) Situating constructionism. In: Harel I, Papert S (eds) Constructionism. Ablex, Norwood, NJ, URL http://www.papert.org/articles/SituatingConstructionism.html

Park J, Choo S (2025) Generative AI prompt engineering for educators: practical strategies. Journal of Special Education Technology 40(3):411–417. https://doi.org/10.1177/01626434241298954

Petri G, Gresse von Wangenheim C (2016) How to evaluate educational games: a systematic literature review. Journal of Universal Computer Science 22(7):992–1021. https://doi.org/10.3217/jucs-022-07-0992

Quiloan R, Moosavi SR, Tankelevich R, et al (2023) PlayGrid: sparking AI literacy with an intuitive and accessible learning interface. In: 2023 Fifth International Conference on Transdisciplinary AI (TransAI). IEEE, pp 171–178, https://doi.org/10.1109/TransAI60598.2023.00045

Resnick M, Maloney J, Monroy-Hernández A, et al (2009) Scratch: programming for all. Communications of the ACM 52(11):60–67. https://doi.org/10.1145/1592761.1592779

Ryan RM, Rigby CS, Przybylski A (2006) The motivational pull of video games: a self-determination theory approach. Motivation and Emotion 30(4):344–360. https://doi.org/10.1007/s11031-006-9051-8

Sailer M, Homner L (2020) The gamification of learning: a meta-analysis. Educational Psychology Review 32(1):77–112. https://doi.org/10.1007/s10648-019-09498-w

Sitzmann T, Ely K, Brown KG, et al (2010) Self-assessment of knowledge: a cognitive learning or affective measure? Academy of Management Learning & Education 9(2):169–191. https://doi.org/10.5465/amle.9.2.zqr169

Tolzin A, Knoth N, Janson A (2024) Worked examples to facilitate the development of prompt engineering skills. In: ECIS 2024 Proceedings, URL https://aisel.aisnet.org/ecis2024/track13_learning_teach/track13_learning_teach/10, paper 10

Touretzky D, Gardner-McCune C, Martin F, et al (2019) Envisioning AI for K-12: what should every child know about AI? In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 9795–9799, https://doi.org/10.1609/aaai.v33i01.33019795

Tsay CHH, Kofinas AK, Trivedi SK, et al (2020) Overcoming the novelty effect in online gamified learning systems: an empirical evaluation of student engagement and performance. Journal of Computer Assisted Learning 36(2):128–146. https:

//doi.org/10.1111/jcal.12385

Walter Y (2024) Embracing the future of artificial intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. International Journal of Educational Technology in Higher Education 21:15. https://doi.org/10.1186/s41239-024-00448-3

White J, Fu Q, Hays S, et al (2023) A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382, https://doi.org/10.48550/arXiv.2302.11382

Woo DJ, Guo K, Susanto H (2025) EFL secondary students' use of ChatGPT for writing task completion pathways. The Journal of Educational Research 118(6):596–609. https://doi.org/10.1080/00220671.2025.2510382

Wood D, Bruner JS, Ross G (1976) The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry 17(2):89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Xiao R, Hou X, Ye R, et al (2025) Improving student-AI interaction through pedagogical prompting: an example in computer science education. arXiv preprint arXiv:2506.19107, https://doi.org/10.48550/arXiv.2506.19107

Zamfirescu-Pereira JD, Wong RY, Hartmann B, et al (2023) Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM, pp 1–21, https://doi.org/10.1145/3544548.3581388