

ChatGPT4PCG Competition: Character-like Level Generation for Science Birds

Pittawat Taveekitworachai*, Febri Abdullah†, Mury F. Dewantoro‡
Ruck Thawonmas§, Julian Togelius§, and Jochen Renz||

*†‡Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan

§College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan

¶NYU Tandon School of Engineering, New York University, New York, USA

||School of Computing, The Australian National University, Canberra, Australia

{*gr0609fv, †gr0397fs, ‡gr0450xi}@ed.ritsumei.ac.jp, §ruck@is.ritsumei.ac.jp, ¶julian@togelius.com, ||jochen.renz@anu.edu.au

Abstract—This paper presents the first ChatGPT4PCG Competition at the 2023 IEEE Conference on Games. The objective of this competition is for participants to create effective prompts for ChatGPT—enabling it to generate Science Birds levels with high stability and character-like qualities—fully using their creativity as well as prompt engineering skills. ChatGPT is a conversational agent developed by OpenAI. Science Birds is selected as the competition platform because designing an Angry Birds-like level is not a trivial task due to the in-game gravity; the quality of the levels is determined by their stability. To lower the entry barrier to the competition, we limit the task to the generation of capitalized English alphabetical characters. We also allow only a single prompt to be used for generating all the characters. Here, the quality of the generated levels is determined by their stability and similarity to the given characters. A sample prompt is provided to participants for their reference. An experiment is conducted to determine the effectiveness of several modified versions of this sample prompt on level stability and similarity by testing them on several characters. To the best of our knowledge, we believe that ChatGPT4PCG is the first competition of its kind and hope to inspire enthusiasm for prompt engineering in procedural content generation.

Index Terms—Angry birds, procedural content generation, large language model, conversational agent, prompt engineering

I. INTRODUCTION

ChatGPT [1] is software that reached 100 million users within two months of its release, making it the fastest application ever to achieve that milestone [2]. It is powered by GPT-3.5, a relatively new version of a large language model (LLM) called Generative Pre-trained Transformer (GPT) developed by OpenAI. ChatGPT is a supervised fine-tuned version of GPT-3.5 using reinforcement learning from human feedback to follow human instructions better [3]. Recent LLMs, including GPT-3.5, are remarkable in size, with GPT-3.5 alone having 175B parameters [3]. Other LLMs [4]–[6], with parameter counts ranging from 280B to 540B, have also been introduced in recent years, enabling unprecedented capabilities not found

in smaller models. Wei et al. [7] discovered that once LLMs reach a certain threshold of parameter count, they can perform tasks without explicit training, exhibiting emergent abilities. These emergent abilities have led to the use of ChatGPT in various use cases [8]–[10].

One notable application of LLMs is in robotics. In a report by Vemprala et al. [11], the authors explore different prompt structures to instruct ChatGPT, a conversational LLM, to output commands for controlling robots in various applications. Procedural content generation (PCG) is another area where LLMs could be useful, especially in level generation. In fact, there already exist studies on applying LLMs for PCG. In Todd et al. [12], they fine tuned variants of GPT-2 to generate levels for Sokoban, a puzzle game, using only the starting token as an input to the model. Similarly, a study by Sudhakaran et al. [13] also applied a fine-tuned version of GPT-2 to generate Mario levels based on instructions. The key difference between these two recent studies is the former focused on exploring the capabilities of LLMs for level generation, but the later focused on controllable level generation using text prompts.

Inspired by these studies, we want to investigate the application of ChatGPT to PCG further. For this purpose, we have chosen Science Birds [14], a clone of Angry Birds, as our evaluation platform because generating levels with good weight distribution that can withstand in-game gravity is challenging. To make it even more challenging, the generated levels must resemble the shape of an English capital letter. To achieve such a level through ChatGPT, one must carefully and creatively design their prompt.

A prompt, in this context, is a text used to communicate with a language model (LM) to get the desired response. The empirical process of designing a prompt to control the behavior of LMs is prompt engineering (PE) [15]. Due to its empirical nature, we have decided to host a competition. The objective of this competition is to find the best prompt that can generate a stable level that looks similar to an English uppercase character. Participants are limited to submitting only a single prompt. Each prompt will be evaluated for all 26

English capital letters. We hope that this competition will encourage people to use both creativity and skills in crafting the prompt that would yield the best results. To do so, we enforce some restrictions for simplicity and provide various tools to aid in and accelerate designing the prompt for the competition.

We hope that this competition will push the boundaries of PE and PCG, in particular, level generation. Through the competition, we anticipate that a better understanding is achieved of PE and PCG with LMs. We also hope to spark an interest in not only interacting but also gaining more understanding of ChatGPT and its emergent abilities. Our contributions are as follows:

- 1) We provide tools and a platform, a modified version of Science Birds, for testing ChatGPT responses in generating levels for Science Birds.
- 2) We provide a sample prompt that serves as an initial point to modify for better-performing prompts.
- 3) We conduct experiments and perform an ablation study on a modified version of our sample prompt to gain a better understanding of its effectiveness and some of the residing components.
- 4) We hope to spark an interest in using ChatGPT for PCG.

We discuss related work in Section II, competition details in Section III, our sample prompt, its modified version as well as variations, and their performance in Section IV, and conclusions in Section V.

II. RELATED WORK

A. Prompt Engineering

The alignment techniques in InstructGPT [3] enable LLMs to produce outputs aligned with users' intentions. This leads to the rise of PE in recent years, with various patterns and techniques being developed to improve the performance of LMs [16], [17]. White et al. [16] identified six categories of prompt patterns, which serve as reusable templates for achieving specific tasks, such as input semantics and output customization. Participants in our ChatGPT4PCG Competition may find these patterns useful for designing or improving their prompts, or they may simply serve as a catalog of available options.

In addition to prompt patterns, there are several notable prompting techniques that have been developed to improve LM performance [18]–[22]. Few-shot prompting [18], [19] provides a few examples of the task in the prompt, allowing for in-context learning and improved performance. Chain-of-thought (CoT) prompting [20] guides the model through a reasoning process with a few examples before reaching a conclusion. Zero-shot CoT prompting [21] does the same thing, but without any examples.

Automatic prompt engineer [22] is an idea where a scoring model automatically generates and selects the best performing prompt. There are more prompt patterns and prompting techniques that have not been mentioned here, and we encourage

participants to experiment and explore some of these techniques feasible for use in the competition, which may result in better performance.

B. Procedural Content Generation

PCG is any automated or semi-automated technique that aims to generate game content using computer software [23]. This content encompasses various objects within a game, such as maps, levels, characters, stories, music, and animation. PCG has a rich history and has been extensively utilized to enhance gameplay mechanics and facilitate game designers and artists in the game design and development process. Notably, a particular aspect of PCG is the level generation, which involves creating levels that meet specific requirements, including visual aesthetics and difficulty levels.

In PCG for level generation, quality and diversity are two critical elements [24]. Quality is the measure of how well the generated levels meet specific criteria, such as playability, difficulty, or playing style. Among these criteria, playability is the most crucial and frequently employed to evaluate the quality of generated levels. On the other hand, diversity determines how different the generated levels are from each other. Insufficient diversity results in levels that are too similar to one another and are often perceived as tedious by players. In contrast, inadequate quality may lead to unplayable levels, which can severely impact the overall gameplay experience.

There already exist studies that use GPT models to generate levels. A study by Todd et al. [12] found in their preliminary results that GPT-3, which has more parameters than GPT-2, requires less data to be fine-tuned. This suggests that the massive training data used in later versions of LLMs makes them more generic for tasks such as PCG. In the context of LLMs for PCG, our work aims to explore the potential of using ChatGPT in PCG as a standalone approach. ChatGPT is capable of generating coherent and contextually relevant text. By employing ChatGPT for game level generation, we aim to contribute to the growing body of research on using natural language for PCG. In this work, we are keen to examine the efficacy of ChatGPT in producing high-quality game levels.

C. Science Birds

Science Birds¹ is a platform developed by Ferreira and Toledo [14] as an extension to the popular game, Angry Birds, with the aim of fostering experimentation in this genre. The platform is developed on Unity game engine. The level's objects all follow the standardized physics rules implemented by the game's engine. Birds can also come in a wide variety, just like blocks that represent the numerous game objects of varied shapes and materials that can be freely used to compose a level structure.

Science Birds has been widely utilized as a test bed for various research endeavors, particularly in the area of level generation [25]–[29]. One of them generates varied, stable, and solvable levels composed of multiple self-contained structures

¹<https://github.com/lucasnfe/science-birds>

[25]. The term “deceptive games” was employed to identify an automated method for creating Angry Birds game levels that are deceptive enough to fool modern AI agents. However, these AI agents may contain pitfalls themselves, making them more vulnerable to deception [26].

To create game levels with more stability and variety, a technique called deep generative model-based level generation was used in one study [27]. This involves encoding the level sequentially as text data, which allows for control of level features through latent variable evolution. In another study [30], a smile interface system was introduced to Science Birds that features Rube Goldberg Machine mechanisms with a domino effect. The system aimed to promote emotions in spectators while they watched the gameplay.

Jiang et al. [29] used pattern-struct and preset-model approaches to create an Angry Birds-like level that resembled a funny quote and won the Fun Track of the CIG 2016 Level Generation Competition. Their approach employed pre-designed sets of blocks that were arranged to form English characters and construct a whole level. This work is similar to ours in generating character-like levels. However, instead of utilizing a pre-defined set of blocks, we challenge participants in our competition to develop general prompts for ChatGPT to generate character-like levels from scratch.

To evaluate generated levels of Science Birds, several evaluation metrics were used in previous studies. The frequently-used evaluation metrics are stability, block frequency, linearity, difficulty, and density [25], [30]. Some studies proposed their PCG approaches in solving specific goals and introduced their own evaluation metrics [30], [31]. For instance, the study by Abdullah et al. [30] aims to generate levels that feature Rube Goldberg Machine mechanisms, and *dynamic* was introduced to measure objects’ movements that indicate the existence of domino effect in the level of interest. However, the most common evaluation metric is stability [25], [27], [30]; stability is important because it closely defines the playability of the level. Hence, we consider stability to be one of our evaluation metrics in this study.

This study also makes use of Science Birds as a platform to visualize and evaluate the efficacy of ChatGPT-generated levels. Specifically, we are interested in evaluating the quality of levels generated by ChatGPT. By leveraging the flexibility and versatility of Science Birds, we aim to gain insights into the effectiveness of this approach and to advance the state of the art in level generation research.

III. CHATGPT4PCG COMPETITION

To simplify the evaluation process in this competition², we introduce constraints on the prompt, requiring participants to develop their prompts that result in ChatGPT generating calls to a specific function named `ab_drop()`, an example of which is shown in Supplementary Material. We also design metrics accordingly to reflect the stability and similarity of each level, which is the main objective of this competition.

²<https://chatgpt4pcg.github.io>

Restrictions and rules regarding the prompt and extensive details on the `ab_drop()` function are discussed in Section III-A. The evaluation process and metrics on stability and similarity are explained in detail in Section III-B.

A. Prompt Rules and Guidelines

In this section, we provide a comprehensive definition and details of the `ab_drop()` function, including expected function behavior, accepted parameters, and the required settings for successful execution of the function. Additionally, we discuss prompt rules on various aspects, such as permissible characters and the maximum number of characters. These rules are intended to ensure fairness and efficiency in the competition.

The `ab_drop()` function is designed to vertically drop a block from the top and center it at a specific slot, denoted by `x_position`. This function shares similarities with the popular game Tetris®. By defining the function in this manner, we aim to reduce the chance of having overlapping blocks in the level. The behavior of the function is simpler than that of placing a block by specifying `x` and `y` coordinates in the level area, which could result in the aforementioned issue. It is worth noting that the `ab_drop()` function is functional in the following settings:

- A structure is situated on a 2D grid with a width (W) of 20 columns and a height (H) of 16 rows. The grid consists of 320 cells, each of equal size. The decision to use a grid of this size is based on the consideration that a larger grid could produce more excessively long output from ChatGPT, consuming a lot of tokens, while a smaller grid may not provide sufficient space for ChatGPT to exercise creativity. Therefore, the chosen size is expected to strike a balance between these two factors.
- Coordinates (x, y) are used to represent the positions in the grid, where x and y show the horizontal and vertical indices of cells, respectively. For example, $(0, 0)$ denotes the bottom-left corner cell of the grid, and $(W-1, H-1)$ is the top-right corner cell.
- A cell on the grid has a size of 1×1 . Each cell has unique (x, y) coordinates associated with it.

This function accepts two parameters, namely, `block_type` and `x_position` with the following details:

- `block_type`: a value that indicates the type of block to be placed. The possible values are `b11`, `b13`, and `b31`, defined in the following and shown in Fig. 1.
 - `b11` denotes a square block whose size is 1×1 .
 - `b13` denotes a column block whose size is 1×3 .
 - `b31` denotes a row block whose size is 3×1 .

This limited selection of block types simplifies the output of ChatGPT and the evaluation process.

- `x_position`: a horizontal index of a grid cell, where 0 represents the leftmost column of the grid, and $W-1$ represents the rightmost column of the grid. The `x_position` parameter indicates the center pivot point of the block being placed. For example, if `b31` is the only block in the level and is placed at `x_position = 4`, it will occupy

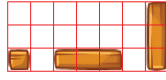


Fig. 1: Block types used in the competition. Starting from the left are b11, b31, and b13, respectively.

cells (3, 0), (4, 0), and (5, 0). An invalid position, like a position where a block of interest intrudes on the grid boundary, will result in an error.

Next, we introduce several restrictions on the prompt, which consist of the following:

- 1) Prompts must be written in English using only alphanumeric characters, space, and the following symbols: `~ / \ + - * ` ' " " " ' . : ; ? - , ! @ # $ % ^ & () _ = [] { } | < > .` This restriction is to ensure that we can better understand them when we select the winning prompt. Furthermore, limiting the prompt to a specific set of allowed characters and symbols helps ensure fairness. This way, no participant can use languages or symbols that others may not have access to.
- 2) The maximum word count for a prompt is 900 words. We arrive at this number heuristically based on the maximum number of tokens accepted by the ChatGPT API, which is 4,096 tokens [32] or approximately 1,800 words. This word limit is important because it ensures that all participants have an equal opportunity to create a prompt without exceeding the maximum number of tokens allowed.
- 3) Prompts designed for this competition must be created for a one-round conversation. This means that each prompt's interaction with ChatGPT should consist of one user request to and one response from ChatGPT. To maintain consistency in our evaluation process, we disallow prompting techniques that require n-round dialogue. This decision is made because LLMs are stochastic by nature, and we cannot guarantee that we will always be able to use all of the prompts provided by each of the participants—who use such techniques—in succession because the maximum number of tokens may be reached before all of them can be used.
- 4) The prompt must include `<OBJECT>` to indicate a section of the prompt that will be replaced by us with each target character, such as “I”, “L”, or “U”. Prompts without `<OBJECT>` will not be assessed.
- 5) To ensure that code blocks can be extracted successfully from responses generated by the ChatGPT API, each output must include three backticks (`````). All content between the last pairs of triple backticks will be extracted to produce a result file containing a sequence of `ab drop()` functions, defined below. It is important to note that the use of loops will be ignored by the script, resulting in only one instance of the loop content being extracted. The use of high-level function approach is chosen based on a study by Vemprala et al. that demonstrated the promising results of generating a sequence

of high-level function calls in various applications.

- 6) The use of ChatGPT plugins³ is not supported, i.e., all plugins are disabled during the evaluation process. This helps ensure fairness since the plugins only available to limited people, a small number of invited developers and web-based ChatGPT Plus users, and ChatGPT API does not support plugins at the time of writing.

B. Competition Evaluation

The submitted prompts will undergo an evaluation process that involves subjecting them to 10 trials for each of the 26 uppercase letters of the English alphabet (A-Z). The decision to conduct 10 trials is made to obtain the average performance of each prompt, which may vary over time due to the non-deterministic nature of ChatGPT. The evaluation of the generated levels for each character will be based on their similarity and stability, which are the focal points of this year's competition. The prompts will be scored using the criteria outlined in the scoring policy, which is provided below. To conduct the entire evaluation process, automated scripts and programs are developed and made available at our GitHub repositories⁴. However, the number of trials and characters in the evaluation set may be adjusted based on the number of participating teams.

1) Evaluation Environment

The evaluation will be conducted within a software and hardware environment as specified below.

- Software
 - OS: Windows 11 Pro
 - Node.js: 18.15.0 LTS
 - Unity: 2019.4.40f1
 - Science Birds Evaluator⁵, a modified version of Science Birds
 - Our automation scripts⁴
- Hardware
 - CPU: Intel(R) Xeon(R) W-2135 CPU @ 3.70GHz
 - RAM: 16 GB

2) Evaluation Process

The evaluation process will be automatically done by scripts as follows:

- 1) The *qualification checking script* will be first run to check for rule violations, including:
 - a) Whether the prompt contains disallowed characters.
 - b) Whether the prompt length exceeds the maximum number of words.
 - c) Whether the prompt does not contain the `<OBJECT>`.
- 2) The *response gathering script* will load each qualified team's prompt and repeat the following two steps for each character.

³<https://openai.com/blog/chatgpt-plugins>

⁴<https://github.com/orgs/chatgpt4pcg/repositories>

⁵<https://github.com/chatgpt4pcg/modified-science-birds>

- a) Replacing `<OBJECT>` with the target character
 - b) Contacting the ChatGPT API for a specific number of trials. Each trial will always start from scratch and will contain no previous conversation.
- 3) The *code extraction script* will load each response and produce a new file containing only a series of `ab_drop()` commands.
 - 4) The *text-to-xml conversion script* will load each code file and convert it into a Science Birds level description XML file. Next, *Science Birds Evaluator*, our modified version of Science Birds, will individually load all levels to assess their stability and capture their images. The results of stability will be recorded, and for each level, an image of the structure with black-textured blocks on a white background will be produced by the program. More details about stability evaluation are given in Section III-B3a.
 - 5) The *similarity checking script* will load each image and pass it through an open-source model called vit-base-letter⁶. It will then record the similarity result. Details about the model and similarity evaluation are described in Section III-B3b.
 - 6) Finally, the *scoring and ranking script* will load all stability and similarity results according to the policy defined in Section III-B3c and produce the final rank and score result for all teams according to the evaluation metrics in III-B3.

3) Evaluation Metrics

Below are our metrics to measure the stability and similarity of generated levels. These metrics serve as evaluation tools and help improve the prompt design. The stability metric evaluates how well the blocks in a level of interest persist after initialization. The similarity metric measures how similar a generated level is to a target character. Using these metrics, prompt designers can improve generation processes for better outcomes. For the rest of the content in this study, i , j , and k represent the indices of trial, target character, and prompt, respectively.

a) Stability

We use *Science Birds Evaluator* to automatically assess stability of levels. We introduce the stability score, st_{ijk} , as follows:

$$st_{ijk} = \frac{total_blocks_{ijk} - moving_blocks_{ijk}}{total_blocks_{ijk}},$$

where $total_blocks_{ijk}$ and $moving_blocks_{ijk}$ are the number of blocks at the initialization step of loading the level into the evaluator and the number of moving blocks during the first 10 seconds after level initialization, respectively. Stable and unstable level examples are shown in Figs. 2a and 2b.

This score represents how well a level can withstand the in-game gravity. A threshold of 0.1 is used for the unit of any blocks moved in Unity, which will allow some room for blocks that are moved by quirks in Unity physics and not caused by

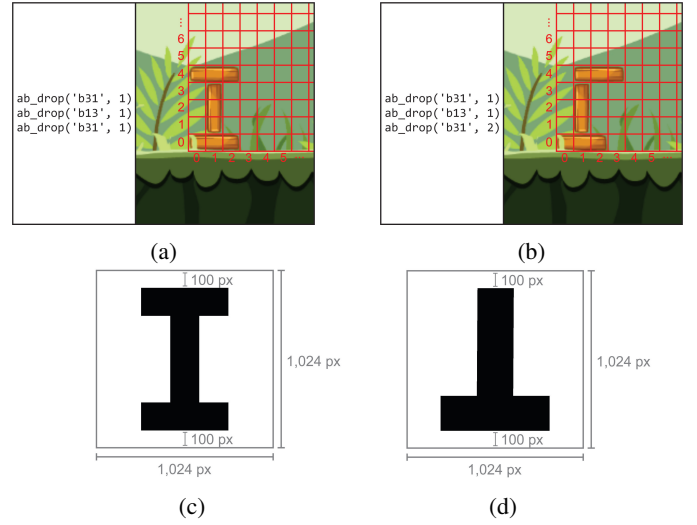


Fig. 2: Examples of stable (a) and unstable (b) levels are shown with their respective remaining block images replaced by black rectangles 10 seconds after initialization (c) and (d).

the structure itself. This metric should encourage participants to create a prompt that can generate a level where most or all of the blocks can withstand the in-game gravity.

b) Similarity

We define the process for evaluating the similarity of a level to the character used in the prompt as the similarity test. The test employs a letter classifier model, described below, and consists of two steps. First, to use the model, an image of the level of interest is required. We utilize *Science Birds Evaluator* to automate the screen capture process. Similarly to the stability test, the screen capture is done 10 seconds after initialization to ensure that the still state of the level of interest is captured—targeting only the remaining blocks and omitting the other details, such as grass and sky. To optimize the capability of the model, the image of each block is replaced by black rectangles, and a white background is used. Captured images always have 1024×1024 pixel resolution with margins of 100 pixels on all sides (Figs. 2c and 2d).

The second step is done by inputting the image of a level into the model to calculate the similarity score, si_{ijk} , as follows:

$$si_{ijk} = \sigma(z_{ijk}),$$

where $\sigma(z_{ijk})$ is the softmax probability of the model in inferring target character j from the image in trial i .

The model is a fine-tuned version of an open-source Vision Transformer (ViT) [33] snapshot, vit-base-patch16-224-in21k⁷ fine-tuned and tested on an open-source letter recognition dataset⁸. The training set comprises 10,000 images for each class, while the testing set consists of 1,000 images per class. The dataset consists of 26 classes, each representing one of the 26 uppercase English

⁶<https://huggingface.co/pittawat/vit-base-letter>

⁷<https://huggingface.co/google/vit-base-patch16-224-in21k>

⁸https://huggingface.co/datasets/pittawat/letter_recognition

characters that the model is trained to predict. The finetuned model achieved an accuracy of 0.9881 and a loss of 0.0515 on the testing set; hyperparameters and other information are available on the online model webpage⁶. Moreover, using a different model for classification can allow for evaluation in diverse domains.

c) Scoring and Ranking

Given that T , C , and P represent the number of trials per target character, the number of characters, and the number of prompts in the competition, respectively, we calculate each prompt's normalized score in the following.

- 1) First, we calculate the weight of each character, $weight_j$, to reflect the difficulty of generating a stable and recognizable shape. This approach ensures that characters that are more challenging to generate are assigned a higher weight, while those that are easier to generate receive a lower weight. This weight is set inversely proportional to the normalized average stability and similarity scores, which are calculated as follows:

$$weight_j = w_{st_j} \times w_{si_j},$$

where

$$w_{st_j} = \max\left(1 - \frac{\sum_{k=1}^P \sum_{i=1}^T st_{ijk}}{PT}, \frac{1}{C}\right)$$

$$w_{si_j} = \max\left(1 - \frac{\sum_{k=1}^P \sum_{i=1}^T si_{ijk}}{PT}, \frac{1}{C}\right)$$

The maximum function, $\max$, sets the weight lower bound to ensure that all characters contribute to scoring calculation, at least to a certain extent.

- 2) Next, the weighted trial score, $trial_{ijk}$, is calculated to determine the performance or target character j in trial i . This score shows how good or bad each trial is when considering both stability and similarity together.

$$trial_{ijk} = weight_j st_{ijk} si_{ijk}$$

- 3) Then, we calculate the average score for character j , $char_{jk}$, as follows:

$$char_{jk} = \frac{\sum_{i=1}^T trial_{ijk}}{T}$$

- 4) For prompt k , its score, $prompt_k$, is calculated by averaging $char_{jk}$ using the following equation.

$$prompt_k = \frac{\sum_{j=1}^C char_{jk}}{C}$$

- 5) Finally, the normalized prompt score, $norm_prompt_k$, is calculated in the following. This score will be used for ranking.

$$norm_prompt_k = 100 \frac{prompt_k}{competition},$$

where

$$competition = \sum_{k=1}^P prompt_k$$

The team that has the highest $norm_prompt_k$ will be declared the winner. If there are multiple teams with the same highest score, the one with the shortest prompt will be chosen as the winner. However, if multiple teams still have the same score and the shortest prompt, they will be considered co-winners.

IV. SAMPLE PROMPT

We provide a sample prompt as a starting point for participants to modify and create their own prompts. The use of this prompt is optional, but participants may find the experiment and results of a modified version of the sample prompt useful in their prompt engineering process. We discuss the structure of the sample prompt in Section IV-A. We explain the experiment details of modified prompts in Section IV-B. Finally, we present results and discuss their implications of the experiment in Section IV-C.

A. Prompt Structure

The sample prompt, shown in Fig. 3, contains four main sections: (1) instructions, (2) definitions, (3) environments, and (4) tools. The first section provides instructions to ChatGPT on what it needs to do, i.e., “[...] to generate a stable structure that looks like the <OBJECT>—the goal.” which helps guide the output of ChatGPT in a way that it contains a series of `ab_drop()` functions. We also provide a marker, i.e., <OBJECT>, to denote a place where *response gathering script* will replace it with the target character like “I”.

Next, we define the construction space where the result of `ab_drop()` will occur, giving ChatGPT a general idea about the constraints of the space. Then we define an environment so that ChatGPT knows more about the possible values to the function, in other words, b11, b31, and b31 blocks used to construct a level. Finally, we specify, again, the use of `ab_drop()` and its definition, so that ChatGPT has more knowledge about how the function works and can use the function to construct the level correctly. These general sections in our prompts are a guideline on what one might include in their prompt. However, we acknowledge that this is not the best possible prompt, and we encourage participants to move, add, or modify the sections to achieve a better outcome.

B. Experiment

The objective here is to assess our modified prompt's performance on some characters to understand its effectiveness. Five variants of the prompt, each with a different set of components, are examined to help us understand the effects of each component on the final outcome. These variants are constructed based on our intuition obtained during the pilot interactions with ChatGPT to develop the sample prompt. The complete prompts for each variant of the modified prompt are available in Supplementary Material. A summary of the variations in each variant, referred to as version (v) from this point, is as follows:

- v1: Detailed explanations about stability have been removed

Use ‘ab_drop()’ function to generate a stable structure that looks like the <OBJECT>—the goal. Dropping position and order are crucial.

1. Definitions

Slots: The map’s width is equally partitioned into W slots where $W = 20$, with slots 0 and 19 being the most left and right, respectively.

Layers: The map’s height is equally partitioned into H layers where $H = 16$, with layers 0 and 15 being the bottom and top layers, respectively. Base: The bottom of the map, i.e., layer 0.

2. Environment

There are three block types as follows:

b11, a square block whose width is 1 unit and height is 1 unit
b31, a horizontal block whose width is 3 units and height is 1 unit

b13, a vertical block whose width is 1 unit and height is 3 units

3. Tool

Use the ‘ab_drop()’ function to vertically drop a block from layer H such that its center is at slot y and drop earlier blocks representing lower parts of the structure.

Fig. 3: A prompt example provided to our participants.

- v2: Any code snippets describing the inner workings of the ab_drop() function have been removed
- v3: A mention of “Tetris®” has been removed
- v4: Any mentions of the “block-stacking problem” have been removed
- v5: No components have been removed.

Three types of English capital characters are used: “I”, “L”, and “U”. These characters were selected to obtain a baseline performance for our experiment due to their low level of difficulty in generating a stable level. Each prompt was run 10 times per character, resulting in a total of 150 levels for evaluation.

Although we plan to use the ChatGPT API in the competition, here we ran all prompts using the free version of the web-based ChatGPT Mar 14 Version and utilized ChatGPT Prompt Genius⁹ to export the prompt result to a Markdown file with a .md extension, one file per trial. If our *code extraction script* failed to extract the response or the *text-to-xml conversion script* failed to convert the level, we re-ran the same prompt using a new conversation session until we obtained a total of 10 trials per character. The remaining evaluation process was conducted using the same process and tools as in the competition, described in Section III-B.

C. Results and Discussions

The results, assessed using the evaluation metrics in Section III-B3, are presented in Table I. The best version, i.e., the one with the highest $norm_prompt_k$, is v1, followed by v4, v5, v3, and v2. An interesting observation is that even

though v4, ranked second, scores the highest for stability in all characters, it does not have the highest $norm_prompt_k$. It is interesting that removing the mention of the “block-stacking problem” from the v4 prompt while retaining the instruction about stability appears to result in better stability performance than vice versa. A code snippet is a useful tool for a model to develop knowledge about new tasks, and removing it significantly reduces performance, as observed with the v2 prompt. Additionally, removing the mention of “Tetris®” as in the v3 prompt also negatively affects performance, because the nature of the game, highly likely presented in the training data, might help the model develop intuition.

The v5 prompt, which consists of all prompt components, ranked third among all versions due to not performing well on the *Similarity* for character “U”, which has the highest w_{st_j} . This is because the evaluation metrics are designed to give more weight to difficult characters, those that a majority of prompts do not perform well in terms of both stability and similarity. It is also suggested that a more detailed prompt may not necessarily yield better performance for this task. For instance, v1 and v4, which are shorter but contain sufficient details, can outperform v5, which is longer.

Table II shows two important observations regarding $weight_j$. Firstly, the similarity weights are all equal since all prompts perform well in stability for all characters, as shown in *Stability* in Table I. Therefore, the first argument of the max function has a very low value, which is then clipped at the lower bound of $\frac{1}{C}$ for each character. Secondly, *Similarity* plays a major role in the final $norm_prompt_k$ score due to having higher weights for all characters. It can be observed that the character “U” has the highest w_{si_j} and $weight_j$, implying that it is the most challenging character to generate a shape that resembles this character, among the three characters in the experiment.

Since the weight of stability is equal, we can focus on only analyzing *Similarity* of version 1, the winner. We found that while it performs moderately well for “I” and “L”, it is the best for “U”. This demonstrates the capabilities of the metrics used in this competition that encourage participants to design a prompt that works well on all characters for both stability and similarity, in particular those that are challenging.

V. CONCLUSIONS AND FUTURE WORK

The ChatGPT4PCG competition has been launched at the 2023 IEEE Conference on Games, aiming to find the best prompt capable of generating both stable and character-like levels from responses of ChatGPT. We provided details on the evaluation process and criteria, a sample prompt, and tools for the competition. In addition, we conducted an experiment to assess the preliminary performance of a modified version, of the sample prompt, and its variants on both stability and similarity using the experimental evaluation set. The results showed that version 1 of our prompt variants had the best performance, which highlights the reason why we designed metrics in a particular way. We hope that this competition—

⁹<https://github.com/benf2004/ChatGPT-Prompt-Genius>

TABLE I: Experiment results where *Stability* and *Similarity* of each variant k denote the average of st_{ijk} and si_{ijk} , respectively, among 10 trials for each character.

	v1			v2			v3			v4			v5		
	I	L	U	I	L	U	I	L	U	I	L	U	I	L	U
<i>Stability</i>	1.00	0.84	0.90	0.96	0.80	1.00	0.97	0.75	0.94	1.00	0.93	1.00	0.96	0.90	1.00
<i>Similarity</i>	0.18	0.53	0.22	0.22	0.24	0.02	0.03	0.65	0.01	0.17	0.60	0.03	0.28	0.43	0.01
<i>norm_prompt_k</i>	30.40			13.26			13.36			23.08			19.90		

TABLE II: $weight_j$ of each character in the experiment.

	I	L	U
w_{st_j}	0.333	0.333	0.333
w_{si_j}	0.823	0.510	0.944
$weight_j$	0.274	0.170	0.315

which we believe is the first of its kind to focus on PCG for PE—will spark interest in both areas.

We believe that ChatGPT or similar tools can generate not only character-like levels but also word-like, object-like, and image-like levels. Furthermore, the use of ChatGPT for PCG should not be limited to level generation but can also extend to other types of content. In the future, we plan to report on the results and findings from this competition, as well as organize similar competitions with different settings to further explore other areas and make them more challenging while also increasing their impact.

REFERENCES

- [1] OpenAI, Nov 2022. [Online]. Available: <https://openai.com/blog/chatgpt>
- [2] K. Hu, “Chatgpt sets record for fastest-growing user base - analyst note,” Feb 2023. [Online]. Available: <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- [3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin *et al.*, “Training language models to follow instructions with human feedback,” 2022.
- [4] J. W. Rae *et al.*, “Scaling language models: Methods, analysis & insights from training gopher,” 2022.
- [5] S. Smith *et al.*, “Using deepspeed and megatron to train megatron-turing nlq 530b, a large-scale generative language model,” 2022.
- [6] A. Chowdhery *et al.*, “Palm: Scaling language modeling with pathways,” 2022.
- [7] J. We *et al.*, “Emergent abilities of large language models,” *Transactions on Machine Learning Research*, 2022, survey Certification. [Online]. Available: <https://openreview.net/forum?id=yzkSU5zdwD>
- [8] A. Gilson *et al.*, “How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment,” *JMIR Med Educ*, vol. 9, p. e45312, Feb 2023. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/36753318>
- [9] E. N. Naumova, “A mistake-find exercise: a teacher’s tool to engage with information innovations, chatgpt, and their analogs,” *Journal of Public Health Policy*, Mar 2023. [Online]. Available: <https://doi.org/10.1057/s41271-023-00400-1>
- [10] M. Dowling and B. Lucey, “Chatgpt for (finance) research: The bananarama conjecture,” *Finance Research Letters*, vol. 53, p. 103662, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1544612323000363>
- [11] S. Vemprala, R. Bonatti, A. Buckner, and A. Kapoor, “Chatgpt for robotics: Design principles and model abilities,” Microsoft, Tech. Rep. MSR-TR-2023-8, February 2023. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilities/>
- [12] G. Todd, S. Earle, M. U. Nasir, M. C. Green, and J. Togelius, “Level generation through large language models,” 2023.
- [13] S. Sudhakaran, M. González-Duque, C. Glanois, M. Freiberger, E. Najarro, and S. Risi, “Mariopt: Open-ended text2level generation through large language models,” 2023.
- [14] L. Ferreira and C. Toledo, “A search-based approach for generating angry birds levels,” in *Proceedings of the 9th IEEE International Conference on Computational Intelligence in Games*, ser. CIG’14, 2014.
- [15] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Comput. Surv.*, vol. 55, no. 9, jan 2023. [Online]. Available: <https://doi.org/10.1145/3560815>
- [16] J. White *et al.*, “A prompt pattern catalog to enhance prompt engineering with chatgpt,” 2023.
- [17] E. Saravia, “Prompt Engineering Guide,” <https://github.com/dair-ai/Prompt-Engineering-Guide>, 12 2022.
- [18] T. B. Brown *et al.*, “Language models are few-shot learners,” 2020.
- [19] S. Min *et al.*, “Rethinking the role of demonstrations: What makes in-context learning work?” 2022.
- [20] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” 2023.
- [21] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” 2023.
- [22] Y. Zhou *et al.*, “Large language models are human-level prompt engineers,” 2023.
- [23] N. Shaker, J. Togelius, and M. J. Nelson, *Procedural content generation in games*. Springer, 2016.
- [24] D. Gravina, A. Khalifa, A. Liapis, J. Togelius, and G. N. Yannakakis, “Procedural content generation through quality diversity,” in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–8.
- [25] M. Stephenson and J. Renz, “Generating varied, stable and solvable levels for angry birds style physics games,” in *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, 2017, pp. 288–295.
- [26] C. Gamage, V. Pinto, J. Renz, and M. Stephenson, “Deceptive level generation for angry birds,” in *2021 IEEE Conference on Games (CoG)*, 2021, pp. 1–8.
- [27] T. Tanabe, K. Fukuchi, J. Sakuma, and Y. Akimoto, “Level generation for angry birds with sequential vae and latent variable evolution,” in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1052–1060. [Online]. Available: <https://doi.org/10.1145/3449639.3459290>
- [28] F. Abdullah, P. Paliyawan, R. Thawonmas, and F. A. Bachtir, “Generating angry birds-like levels with domino effects using constrained novelty search,” in *2020 IEEE Conference on Games (CoG)*. IEEE, 2020, pp. 698–701.
- [29] Y. Jiang, T. Harada, and R. Thawonmas, “Procedural generation of angry birds fun levels using pattern-struct and preset-model,” in *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, 2017, pp. 154–161.
- [30] F. Abdullah, P. Paliyawan, R. Thawonmas, T. Harada, and F. A. Bachtir, “An angry birds level generator with rube goldberg machine mechanisms,” in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–8.
- [31] M. Kaidan, T. Harada, C. Y. Chu, and R. Thawonmas, “Procedural generation of angry birds levels with adjustable difficulty,” in *2016 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2016, pp. 1311–1316.
- [32] OpenAI. [Online]. Available: <https://platform.openai.com/docs/models/gpt-3-5>
- [33] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.