Evaluating the Pokémon Generator

Leiden University, Media Technology MSc
Computational Creativity

Christina Irakleous (s3729664) Nikki Rademaker (s2641887)
Yanna Smid (s2634163) Generative AI (AI2023GPT3)

December, 2023

1 Introduction

Over the years the Pokémon franchise has evolved incredibly. Starting with a black-and-white pixel video game in the late 90s it quickly transformed into a worldwide famous phenomenon with games, trading cards, and animated series. The popularity mainly revolved around the well-designed pocket monsters, known as Pokémon, that the world grew attached to. Over the generations of Pokémon, many unique Pokémon designs were developed. Catching fans' interests, and inspiring fans to create their own. Now with the current advanced technology, it is possible to let AI come up with new Pokémon, based on a person's preferences. However, no practical system exists yet that allows a person to design a new Pokémon with the cooperation of AI. Therefore, a co-creative system has been implemented that ensures cooperative development between a human and a computer. This report documents the implementation of the co-creative system and reflects on it. Furthermore, it discusses two different user studies that were done to evaluate the Pokémon generator and its output.

1.1 Implementation and Reflections

The AI Pokémon generator was made with Python. It is a co-creative system that combines user input with generative AI. The system shows an interface that allows the user to choose two Pokémon types, the body type, and a short description of what the Pokémon should be inspired by. The user's input is then used in the prompt for the AI to generate the text with the description of the new Pokémon. The created text is then used to generate an image of the Pokémon. After the AI is done with generating the new Pokémon, it displays the text and the image on the interface screen and the user gets to see the Pokémon they just have created together with the system.
The AI Pokémon Generator system relies on the functions provided by the OpenAI API. Before the available functions from the Openai library could be used, certain basic steps had to be undertaken first. The functions from the Openai package could only be used if either usage credits were bought or a paid membership account was used. Furthermore, after the payments are finalized, an API key needs to be created with the same account. This API key needs to be copied into a dotenv file, so a link can be created between the system's code and the OpenAI account.
To generate the text that describes the new Pokémon, the *gpt-3.5-turbo* model is used. The prompt is automatically filled in based on the user's input to ensure that the Pokémon meets the user's

selections. However, in the early stages of the system's development, the generative tool struggled to stick to the word limit mentioned in the prompt. For this reason, the max tokens parameter was set to 1024 instead of 2048. Consequently, the provided text was too long to use for the image generator function, that used the *dall-e-3* model. Therefore, only the first two paragraphs of the AI Pokémon description were used in the prompt for the image generator.
In the end, the generative AI system is a major contribution to the co-creative system implemented for this assignment. The functions of the Openai library were extremely helpful and allowed the user to cooperate with an AI to design their own Pokémon.

2 User Evaluations

2.1 Turing Test

In the evaluation of the developed Pokémon generator, the Turing Test was applied as a measure of the ability of the system to mimic human-like creativity. This test, originally conceived by Alan Turing, assesses whether a machine's output is indistinguishable from a human's.
In this case, it involved distinguishing between Pokémon images generated by AI and those created by humans. Twenty-seven participants, with varying levels of Pokémon knowledge, were presented with a series of thirteen Pokémon images through a Google Forms questionnaire. They were tasked with identifying each image as either a Pokémon created by AI or fan art of an existing Pokémon made by a human. The knowledge of the participants influences their ability to distinguish between AI-generate and human-created Pokémon. Therefore, in addition, the participants had to rate their familiarity with Pokémon on a scale from 1 (not familiar) to 5 (very familiar). The aim of this test was to see if the AI could convincingly replicate the artistic style of real Pokémon.

The images used for the Turing Test, and the results of the Turing Test can be found in the Appendix A.1. Eight Pokémon (Figures 3, 9, 11, 15, 17, 19, 23, and 27) were generated by the Pokémon generator, meaning they were generated by AI. Five Pokémon (Figures 5, 7, 13, 21, and 25) were human-made.
Analyzing the Turing Test results for your Pokémon generator reveals that it is able to effectively mimic the artistic style of real Pokémon. The test was structured as a guessing game, with a 50% chance of participants making a correct identification. It turned out, in 10 out of the 13 cases (Pokémon 1, 2, 3, 5, 6, 7, 9, 10, 12, and 13), fewer than half of the participants could correctly classify whether the Pokémon were AI-generated or human-made. This outcome, particularly for AI-generated Pokémon 1, 5, 7, 9, and 13, and human-made Pokémon 2, 3, 6, 10, and 12, suggests that the generator successfully blurred the lines between AI and human creativity. In contrast, Pokémon 4, 8, and 11, all AI-generated, were more easily identified by participants, with approximately 85%, 67%, and 56% accuracy, respectively. These Pokémon did not pass the Turing Test. This shows that in some cases the AI's creative capabilities were more distinguishable from human-made Pokémon. Most of the respondents were unfamiliar or somewhat familiar with Pokémon. Only four people considered themselves to be familiar with Pokémon. However, out of these four participants, just two were noticeable better at correctly classifying the Pokémon.
Overall, the test results strongly suggest that the Pokémon generator can convincingly replicate the style of real Pokémon. Most of the images challenged the respondents' ability to differentiate between human and AI creations, highlighting the advanced creative capabilities of the AI system.

Using AI for the evaluation provided valuable insights. It provided a strong indication that co-creative system is able to mimic human creativity and create new Pokémon that are difficult to distinguish from existing Pokémon.

2.2 Evaluating Creativity

To further evaluate the creativity of the Pokémon generator, the Standardized Procedure for Evaluating Creative Systems (SPECS) of Jordanous was used. For the Pokémon generator, creativity is defined as the system's ability to produce unique Pokémon that are coherent and appealing, combining elements from the selected criteria in the style of Pokémon. Based on this definition, the following criteria are formulated to evaluate the creative system:

1. **The cohesiveness of Elements**: the Pokémon's design should maintain a cohesive appearance where the types, body type, and inspired look form a complete and finished creature.
2. **Thematic Consistency**: The design should reflect the themes associated with the chosen types and inspired look.
3. **Feasibility within the Pokémon Universe**: The design should be plausible within the context of the Pokémon, meaning it should not be overly complex or unrealistic to the point where it couldn't be imagined as part of the Pokémon universe.
4. **Originality of Design**: The generated Pokémon should be original and not look like any of the already existing Pokémon.
5. **Descriptive Clarity**: The text should clearly describe the unique features of the generated Pokémon, including its name, appearance, abilities, and any distinctive traits linked to its types and inspired look.
6. **Intention and emotional involvement**: The Pokémon should be inspiring to the user and create some kind of emotional attachment to it.

To investigate if the system meets these requirements, 10 questions needed to be answered by the user after they had used the co-creative system. Each criterion is supported by questions that the user needs to answer with a rating on a scale from 1 to 5. An overview of the questions corresponding to the criteria is given in the appendix A.2. The average scores are compared to the aimed score that was set for each question. The table in figure 1 summarizes these scores. According to the results, the co-creative system scores high on the first and fifth criteria. This means that overall the generated Pokémon are cohesive and complete, and the generated text is sufficient and clear. Furthermore, the average score of the second and fourth criteria also peaked above the target score, meaning that the generated Pokémon are a good representation of the user's input, and original in their design. For the last criterion, it appears that users experience emotional value and inspiration from the Pokémon very differently. Some users felt ultimately inspired and attached to the Pokémon, whereas other users gave a neutral score of 2 or 3. Overall, on average the emotional attachment requirement meets the target score, but in regards to inspiration, the system still misses one-tenth of a score. The system scores below the target value when it comes to the feasibility of the Pokémon within the Pokémon Universe. This means that, together with the generative AI, the creative system needs to work on the imaging of Pokémon that fit the right

style. More time needs to be spent to optimize the prompt that is used to generate the images. Furthermore, the OpenAI model needs to be improved to create images more strictly as instructed. Even though the target value for this criterion is not reached, the average score is still above the worst acceptable outcome. In conclusion, all criteria meet the minimum acceptable score, indicating that the co-creative AI Pokémon generator can be classified as creative according to the previously mentioned definition.

SPECS	statements	average score	target score	worst possible score	worst acceptable score	best possible score
Cohesiveness of Elements	cohesive appearance	4.8	4	1	3	5
Thematic Consistency	image represents user's input	4.2	4	1	3	5
Feasibility within the Pokémon Universe	pokemon is atypical and promising	3.8	4	1	3	5
	pokemon fits in pokemon universe	3.6	4	1	3	5
Originality of Design	unique image and description	4.3	4	1	3	5
	unique design and concept	4.3	4	1	3	5
Descriptive Clarity	text reflects user's input	4.4	4	1	3	5
	accurate text description	4.6	4	1	3	5
Intention and emotional involvement	feeling emotional attachment	4	4	1	3	5
	feeling inspired	3.9	4	1	3	5

Figure 1: Table containing all the questionnaire statements and their results for each set criterion of the SPECS. The average score column shows the average scores for each statement determined over the 17 different participants. The target score column shows the average scores that were aimed at. The worst possible score column shows the worst possible scores that could have been given for each statement. The worst acceptable score column shows the worst scores that would still be considered as acceptable for the creativity of the co-creative system. The best possible score column shows the best scores that were possible to get. Almost all statements are the same or above the set target scores, except for "I think the generated Pokémon was pretty atypical but also highly valued and promising" and "I felt inspired by the generated Pokémon", where the average scores were approximately 3.6 and 3.9, respectively.

3 Future Work

The co-creative AI Pokémon generator is far from finished. The system includes a simple interface with still a few impractical functions. To improve the system, an option should be added to save

the generated image and copy the provided description. This enhances the users' experience when creating a new Pokémon with the generative AI. Furthermore, as discussed in section 2.2, the generative AI needs to be adjusted to increase the satisfaction of the third criterion, regarding the Pokémon's feasibility. This guarantees a better image output, that fits the user's preferences. Lastly, the system can be further developed to publish it on an open platform to make it available to every Pokémon fan.

References

[1] Thievul by Pixiv ID 5954606. Inesanemona. zerochan (November, 2019). https://www.zerochan.net/2763425

[2] Meowscarada by Kampidh. Krita Artists (January, 2023). https://krita-artists.org/t/meowscarada/56401

[3] Inteleon by gino_beber_art. ArtStation (2022). https://www.artstation.com/artwork/KeeYZB

[4] Spectrier by Cynthiaramirez123. Zhu Zhu Pets Fanon Wiki. https://zhu-zhu-pets-fanon.fandom.com/wiki/Spectrier_(Pok%C3%A9mon)

[5] Grafaiai by Andy (Pandabox). ArtStation (March, 2023). https://www.artstation.com/artwork/KeeYZB

A Appendix

A.1 Turing Test Results

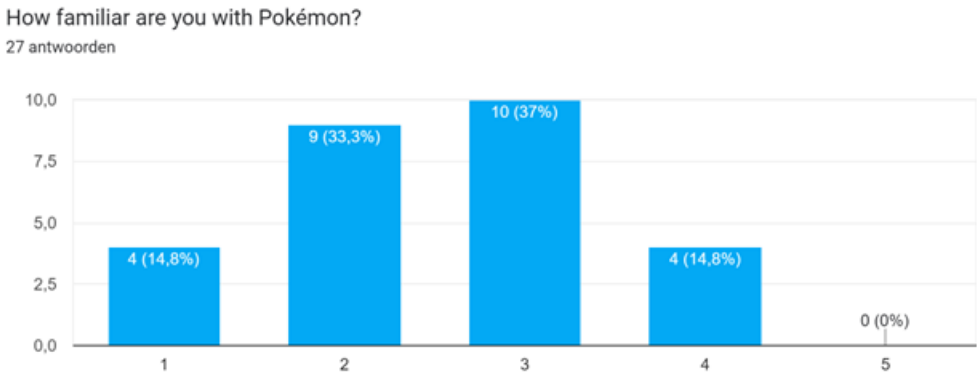

Figure 2: Response of how familiar the participants of the Turing Test were with Pokémon.

Figure 3: The first Pokémon showed during the Turing Test. This Pokémon was created by the Pokémon generator, meaning it was generated by AI.

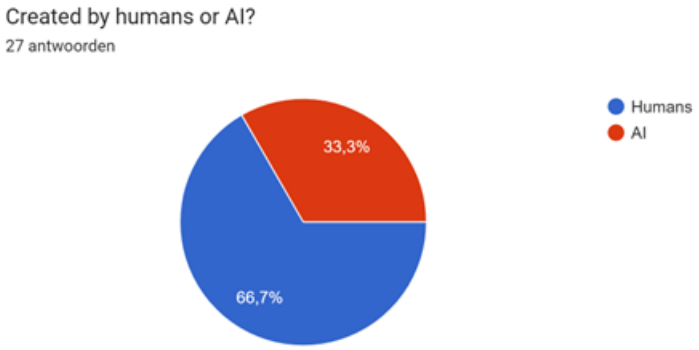

Figure 4: Results of the Turing Test for the first Pokémon. This Pokémon was generated by AI. Approximately 33% of the participants correctly answered this question.

Figure 5: The second Pokémon showed during the Turing Test. This Pokémon was created by humans [1].

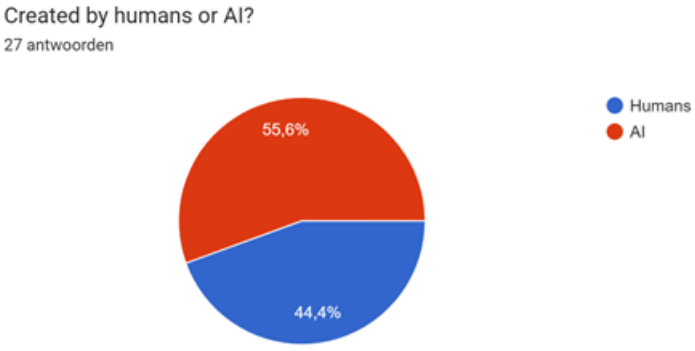

Figure 6: Results of the Turing Test for the second Pokémon. This Pokémon was generated by humans. Approximately 44% of the participants correctly answered this question.

Figure 7: The third Pokémon showed during the Turing Test. This Pokémon was created by humans [2].

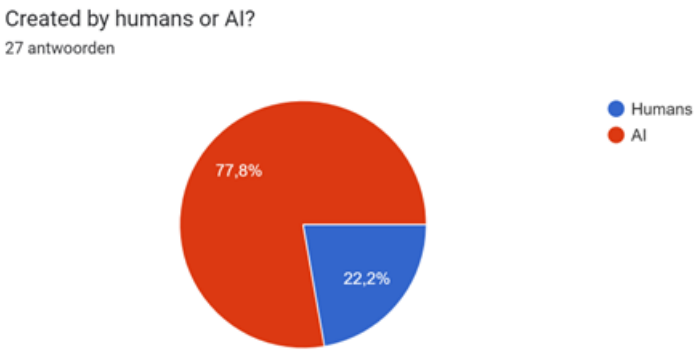

Figure 8: Results of the Turing Test for the third Pokémon. This Pokémon was generated by humans. Approximately 22% of the participants correctly answered this question.

Figure 9: The fourth Pokémon showed during the Turing Test. This Pokémon was created by AI.

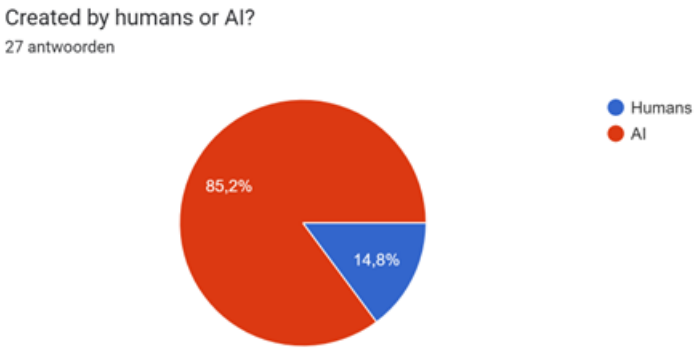

Figure 10: Results of the Turing Test for the fourth Pokémon. This Pokémon was generated by AI. Approximately 85% of the participants correctly answered this question.

Figure 11: The fifth Pokémon showed during the Turing Test. This Pokémon was created by AI.

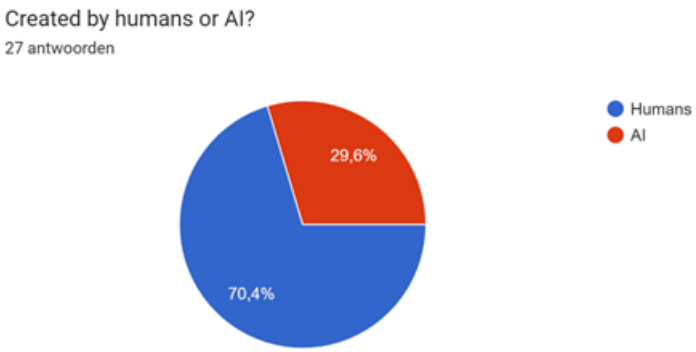

Figure 12: Results of the Turing Test for the fifth Pokémon. This Pokémon was generated by AI. Approximately 30% of the participants correctly answered this question.

Figure 13: The sixth Pokémon showed during the Turing Test. This Pokémon was created by humans [3].

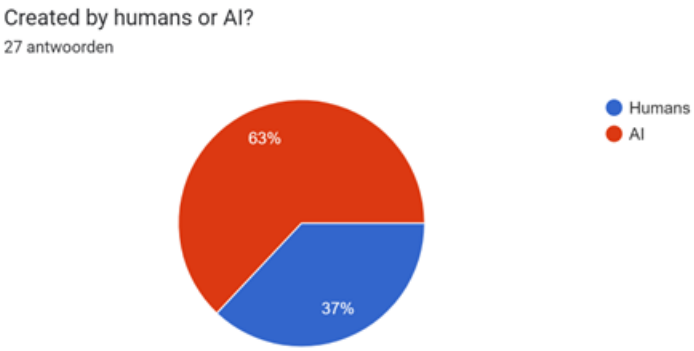

Figure 14: Results of the Turing Test for the sixth Pokémon. This Pokémon was generated by humans. Approximately 37% of the participants correctly answered this question.

Figure 15: The seventh Pokémon showed during the Turing Test. This Pokemon was created by AI.

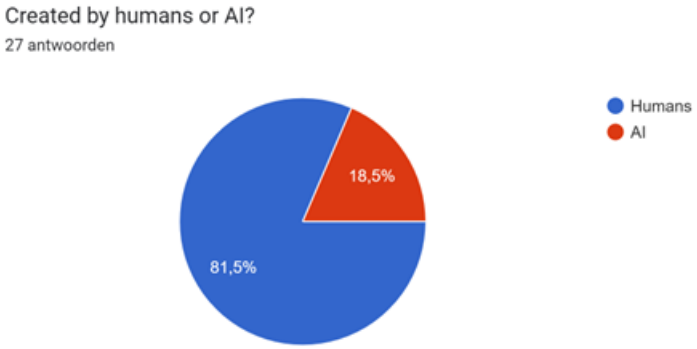

Figure 16: Results of the Turing Test for the seventh Pokémon. This Pokémon was generated by AI. Approximately 19% of the participants correctly answered this question.

Figure 17: The eighth Pokémon showed during the Turing Test. This Pokémon was created by AI.

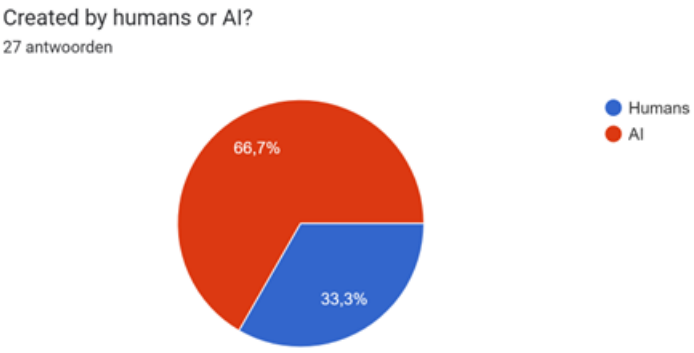

Figure 18: Results of the Turing Test for the eighth Pokémon. This Pokémon was generated by AI. Approximately 67% of the participants correctly answered this question.

Figure 19: The ninth Pokémon showed during the Turing Test. This Pokémon was created by AI.

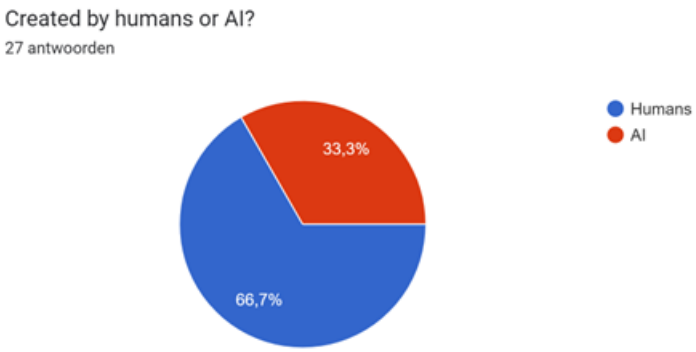

Figure 20: Results of the Turing Test for the ninth Pokémon. This Pokémon was generated by AI. Approximately 33% of the participants correctly answered this question.

Figure 21: The tenth Pokémon showed during the Turing Test. This Pokémon was created by humans [4].

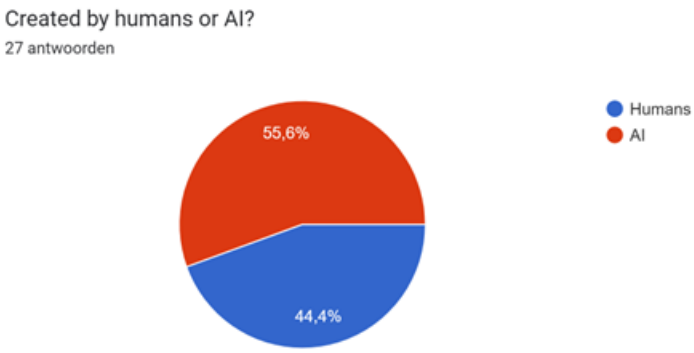

Figure 22: Results of the Turing Test for the tenth Pokémon. This Pokémon was generated by humans. Approximately 44% of the participants correctly answered this question.

Figure 23: The eleventh Pokémon showed during the Turing Test. This Pokémon was created by AI.

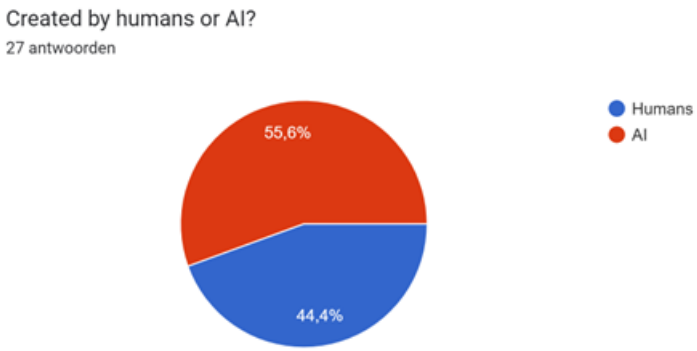

Figure 24: Results of the Turing Test for the eleventh Pokémon. This Pokémon was generated by AI. Approximately 56% of the participants correctly answered this question.

Figure 25: The twelfth Pokémon showed during the Turing Test. This Pokémon was created by humans [5].

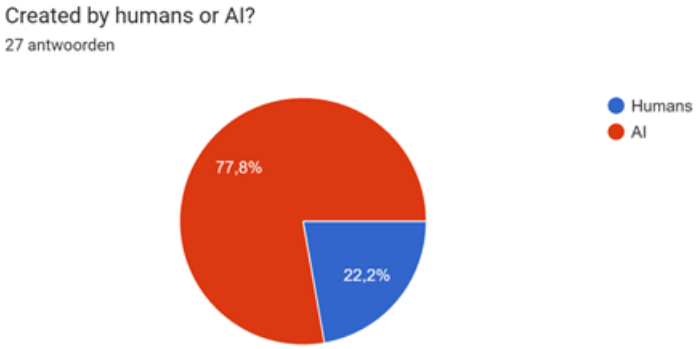

Figure 26: Results of the Turing Test for the twelfth Pokémon. This Pokémon was generated by humans. Approximately 22% of the participants correctly answered this question.

Figure 27: The thirteenth Pokémon showed during the Turing Test. This Pokémon was created by AI.

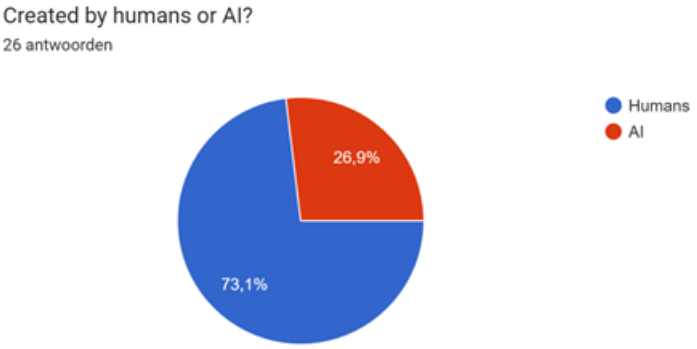

Figure 28: Results of the Turing Test for the thirteenth Pokémon. This Pokémon was generated by AI. Approximately 27% of the participants correctly answered this question.

A.2 SPECS Results

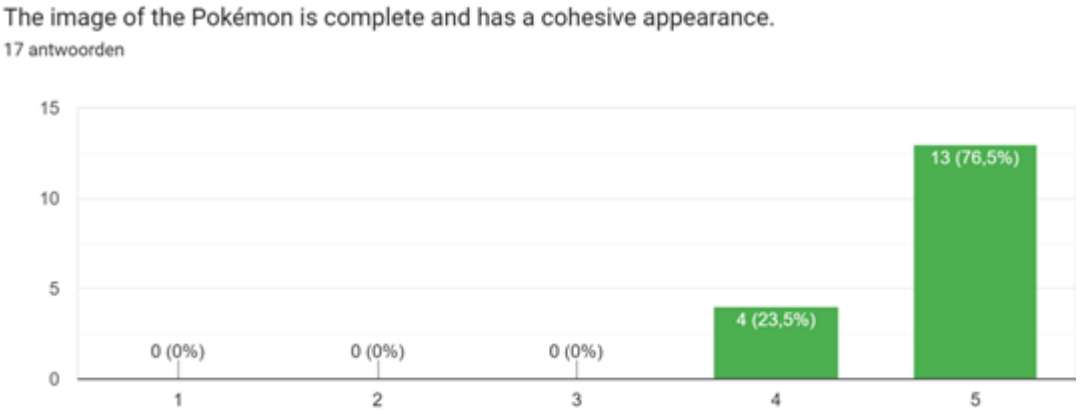

Figure 29: Results of the question aimed at investigating the first criterion regarding the cohesiveness of the generated Pokémon.

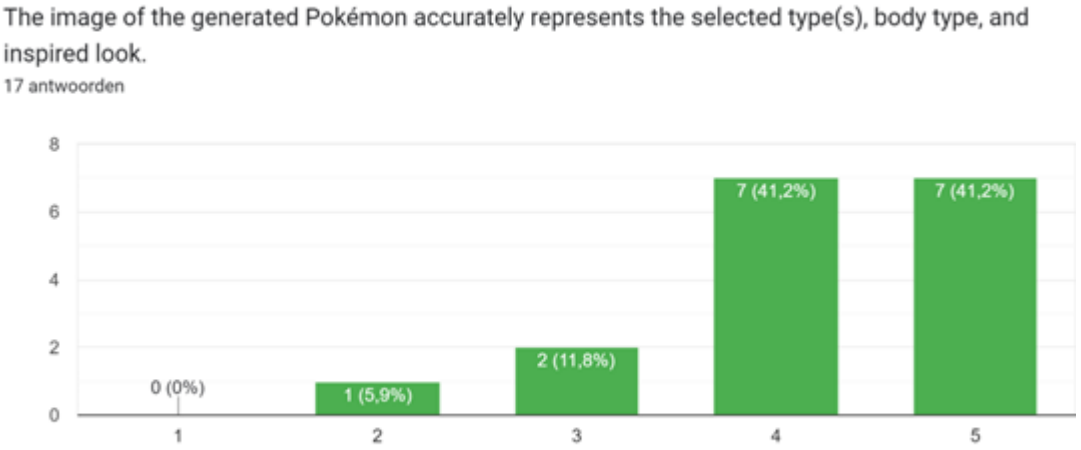

Figure 30: Results of the question aimed at investigating the second criterion regarding the thematic consistency of the AI Pokémon.

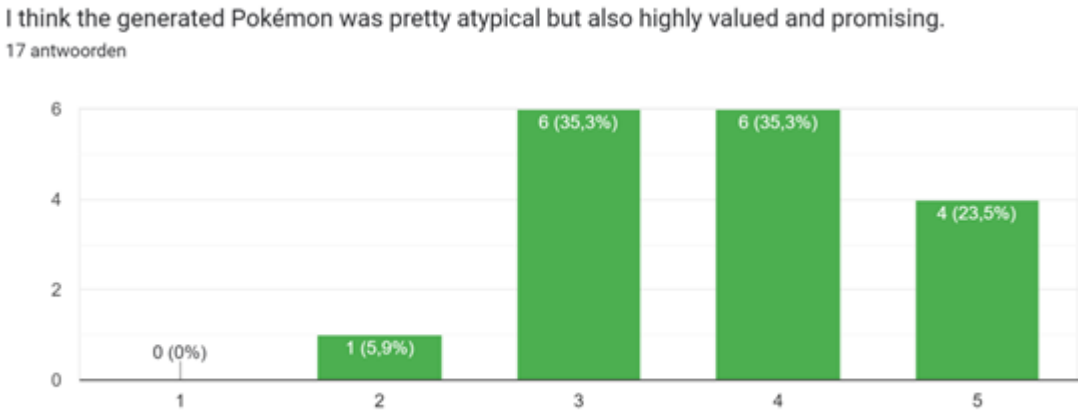

Figure 31: Results of the first question aimed at investigating the third criterion regarding the feasibility of the AI Pokémon.

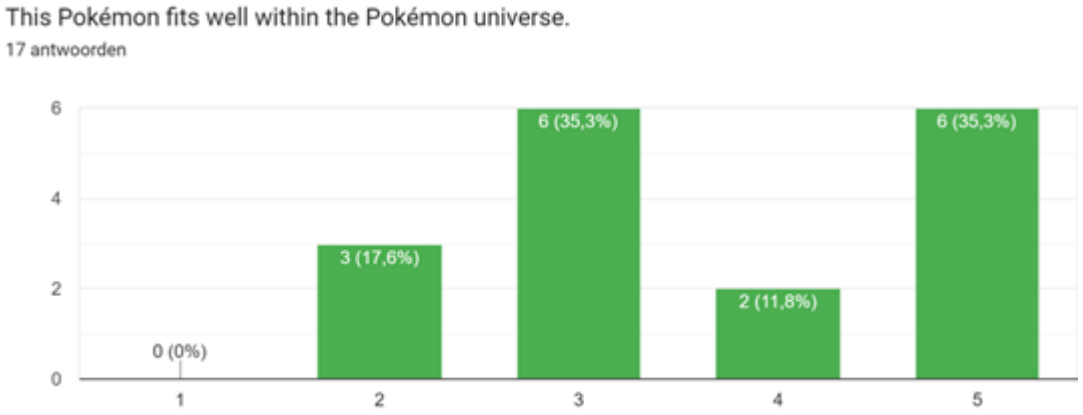

Figure 32: Results of the second question aimed at investigating the third criterion regarding the feasibility of the AI Pokémon.

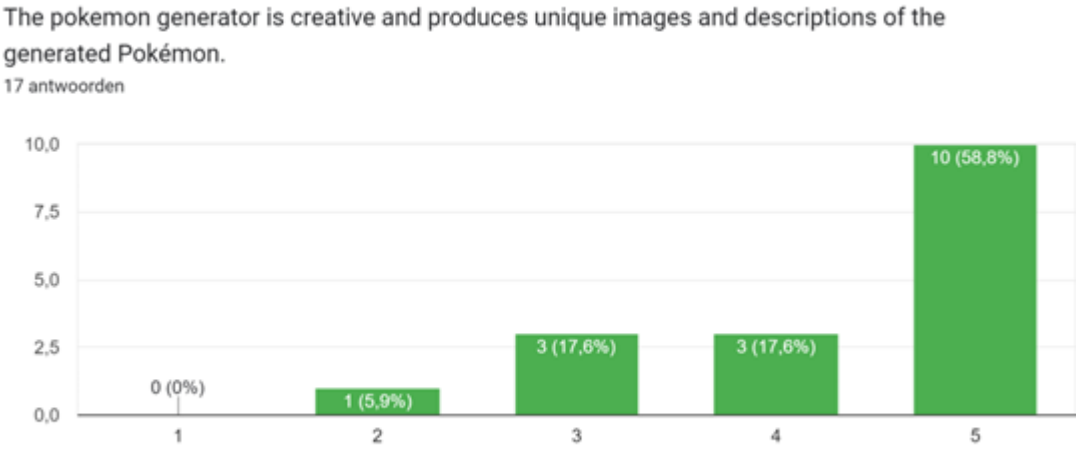

Figure 33: Results of the first question aimed at investigating the fourth criterion regarding the originality of the AI Pokémon.

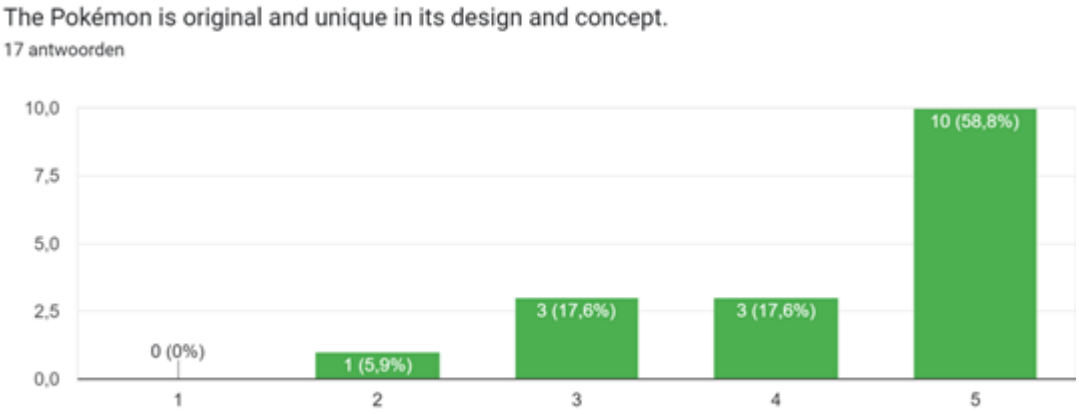

Figure 34: Results of the second question aimed at investigating the fourth criterion regarding the originality of the AI Pokémon.

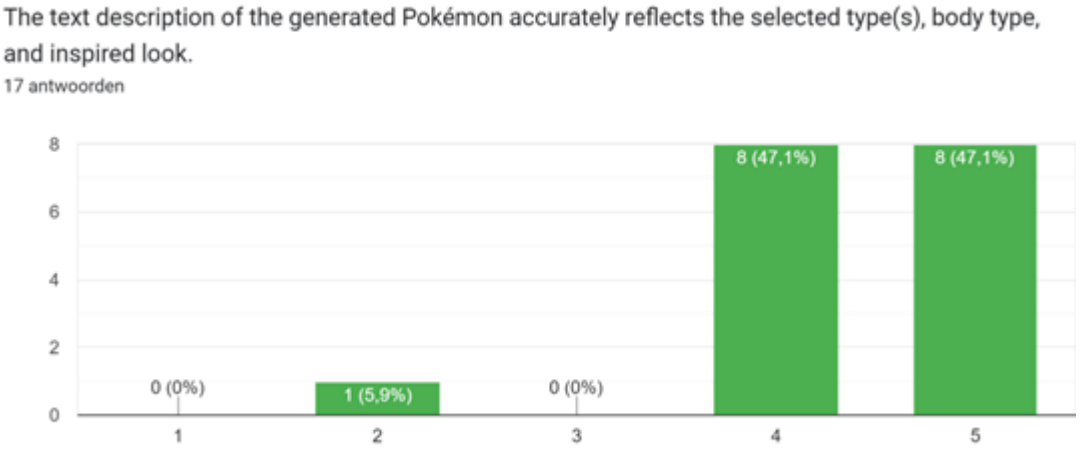

Figure 35: Results of the first question aimed at investigating the fifth criterion regarding the clarity of the text-based description of the AI Pokémon.

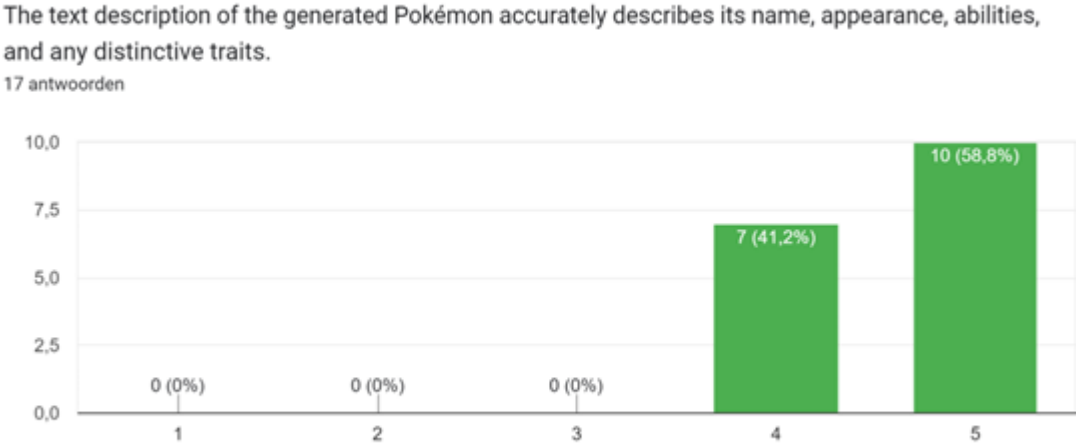

Figure 36: Results of the second question aimed at investigating the fifth criterion regarding the clarity of the text-based description of the AI Pokémon.

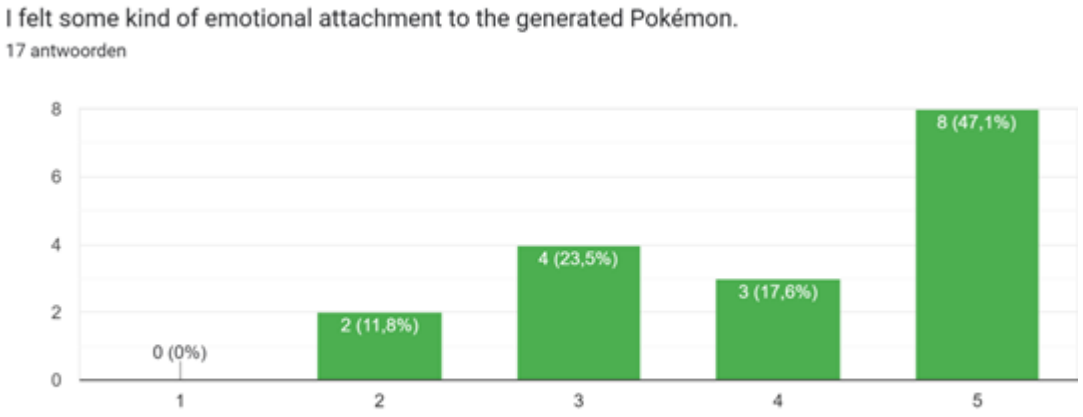

Figure 37: Results of the first question aimed at investigating the sixth criterion regarding the clarity of the text-based description of the AI Pokémon.

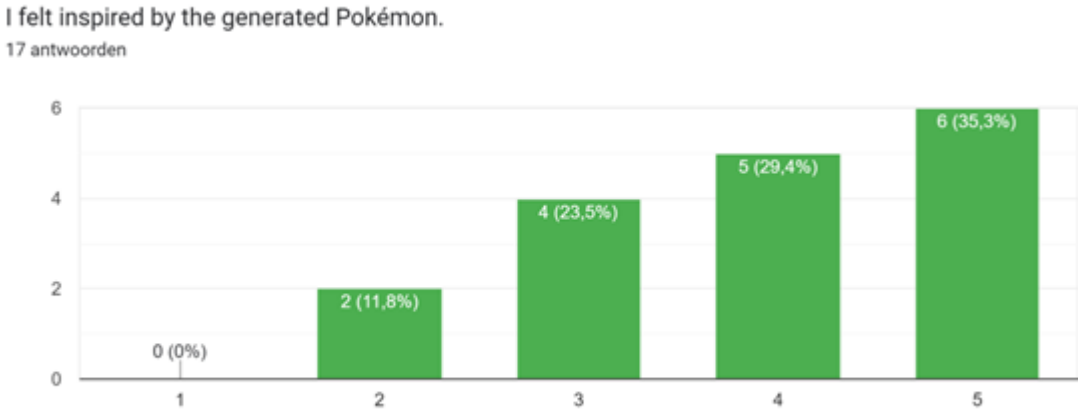

Figure 38: Results of the second question aimed at investigating the sixth criterion regarding the clarity of the text-based description of the AI Pokémon.