USING AI FOR TEST ITEM GENERATION:

opportunities and challenges

OLENA ROSSI, PhD
olena.rossi@itemwriting.co
itemwriting.co
@olena_rossi

The rapid expansion of technology-mediated language assessment (TMLA)

Exponentially growing demand for test items

Producing items by item writers is time-consuming and expensive

Lack of trained item writers who can produce high-quality items

Rapid development of AI technology

↓

USING AI FOR LANGUAGE TEST ITEM GENERATION

- Automated item generation and item generation with AI
- Large language models and how they work
- GPT: opportunities and challenges
- Using ChatGPT for test item generation in classroom assessment
- Using GPT for item generation in large-scale testing (DET)
- Using large language models for language test item generation: future directions

Automated item generation

Step 1: item model development by item writers (prototype, template)

Step 2: Identifying content to be manipulated

Step 3: Generating items using computer technology

(Gierl & Lai, 2013)

Using AIG elements to produce TOEIC listening items:

an automated system that retrieves appropriate real-world videos

a list of vocabulary tagged with established difficulty levels

a word similarity tool that suggests words and phrases similar in distribution to a given word

(Yoon et al, 2017)

Item generation with AI

ITEM WRITING for language testing

Natural Language Processing (NLP) (ChatGPT, accessed on 24 May 2023)

Large language models such as BERT and GPT

ChatGPT is freely available https://chat.openai.com

- based on GPT3.5 version
- ChatGPT+ (paid version) based on GPT4

Most recent model – GPT4 (released on 14th March 2023 by OpenAI):

- solves complex problems with greater accuracy
- better understands nuances of language including the meaning, tone and context
- can generate longer texts and texts of higher quality
- can better remember the context of a conversation for longer (BingChat, accessed on 24 May 2023)

Bing Chat (freely available as part of Microsoft Edge browser)

- does not have a knowledge-cut-off
- provides links to webpages with further information

How do GPT large language models work?

trained on a massive dataset of human-generated texts

utilizes advanced NLP techniques to understand and generate language

learns to generate text by analyzing and mimicking the text it is trained on

produces fully original texts (0% similarity to online sources) by predicting the next word (token) in a sequence

can produce texts of any genre, on any topic, in a matter of seconds

the texts are written in correct English and are largely indistinguishable from texts written by humans (Christodoulou, 2023)

OPPORTUNITIES

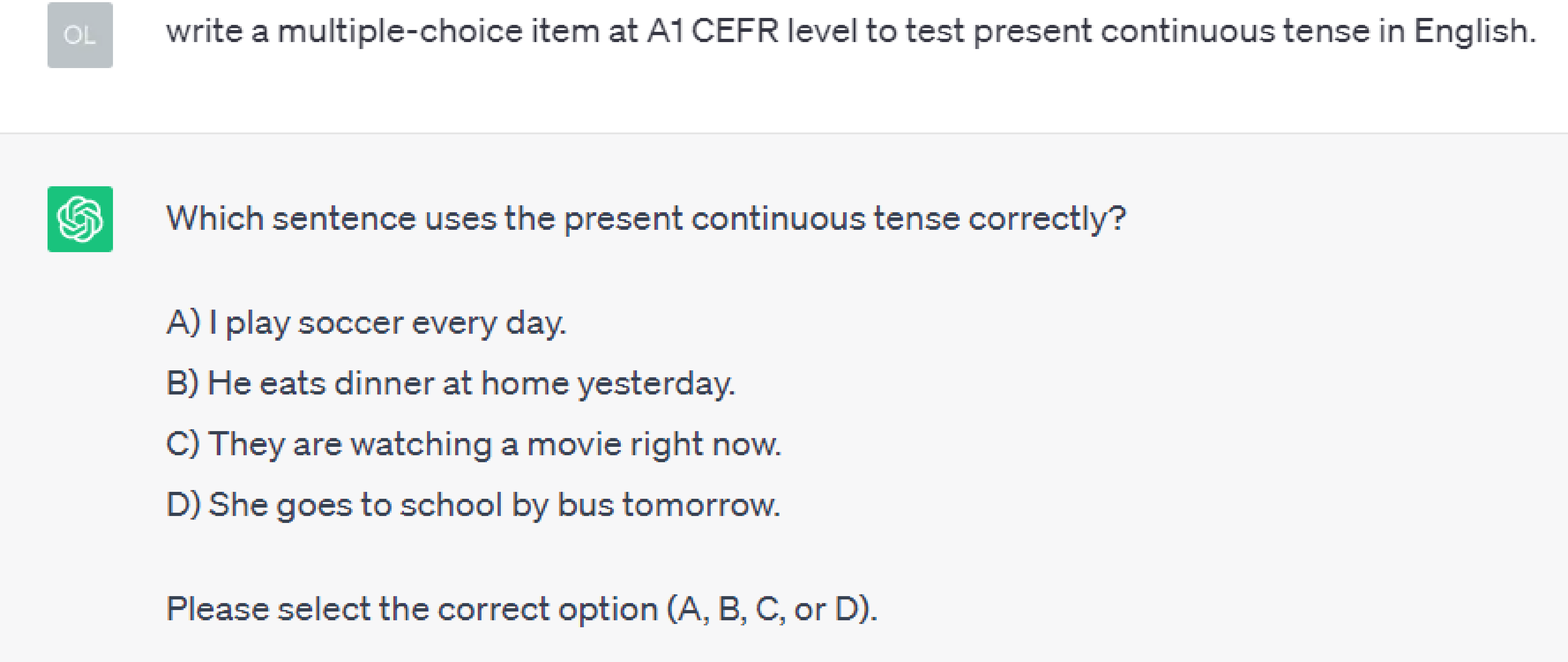
- GPT generates texts of any genre on any topic in seconds
- Latest GPT models can produce output with little to no demonstration (zero-shot learning)
- texts are written in good English
- GPT can be fine-tuned to produce better output (Open AI’s API)

CHALLENGES

- GPT ‘hallucinates’
  - GPT *“was prone to generating erroneous outputs…, especially for certain statistics and numbers”* (Bezirhan & von Davier, 2023, p.17)
- GPT is biased
  - *biased gender representation* (Li & Bamman, 2021; Shihadeh et al., 2022)
  - *racial bias* (Brown et al., 2020; Log’e et al., 2021)
  - *religious bias* (Abid et al., 2021*).*
- Lack of clarity regarding text authorship (Zong & Krishnamachari, 2022)

Using ChatGPT for test item generation: Grammar MCQs

ITEM WRITING
for language testing



write a multiple-choice item at A1 CEFR level to test present continuous tense in English.

Which sentence uses the present continuous tense correctly?

A) I play soccer every day.

B) He eats dinner at home yesterday.

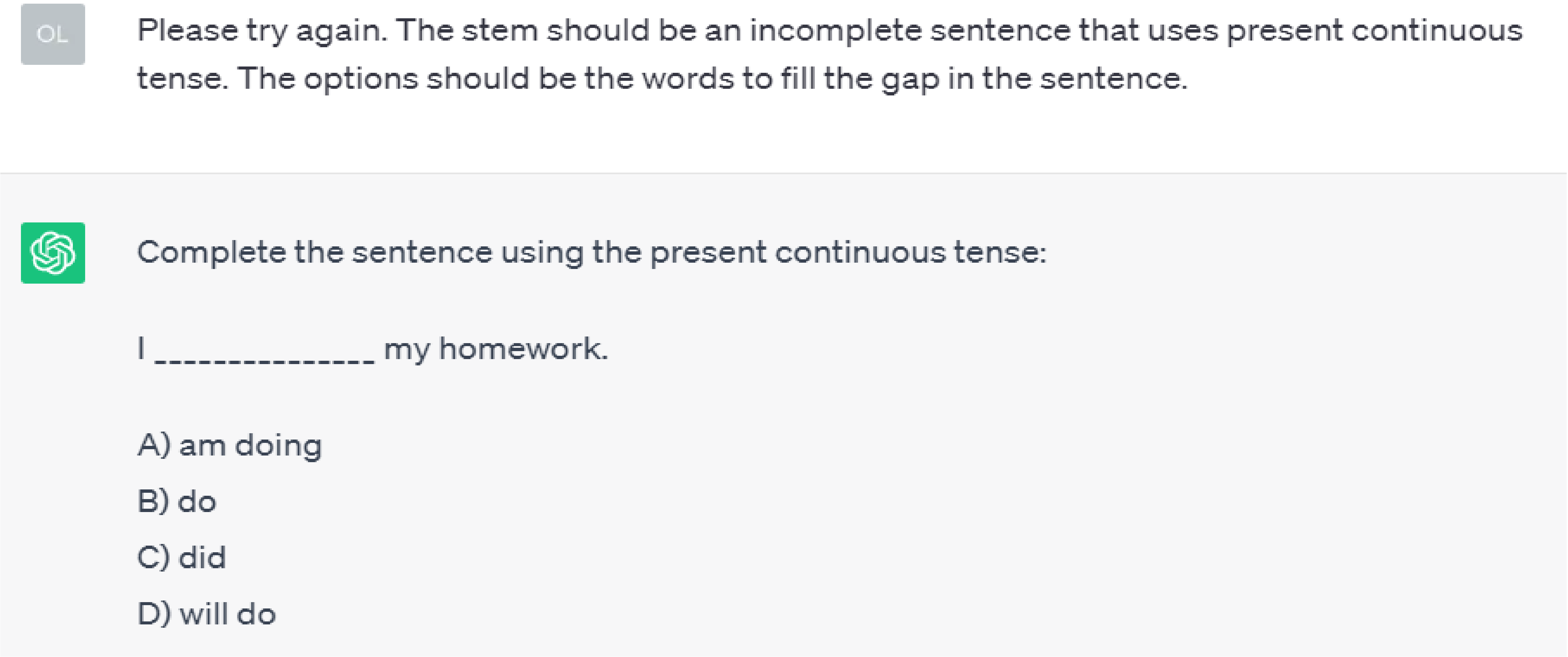
C) They are watching a movie right now.

D) She goes to school by bus tomorrow.

Please select the correct option (A, B, C, or D).

ChatGPT, accessed on 25th May 2023

Using ChatGPT for test item generation: Grammar MCQs



Please try again. The stem should be an incomplete sentence that uses present continuous tense. The options should be the words to fill the gap in the sentence.

Complete the sentence using the present continuous tense:

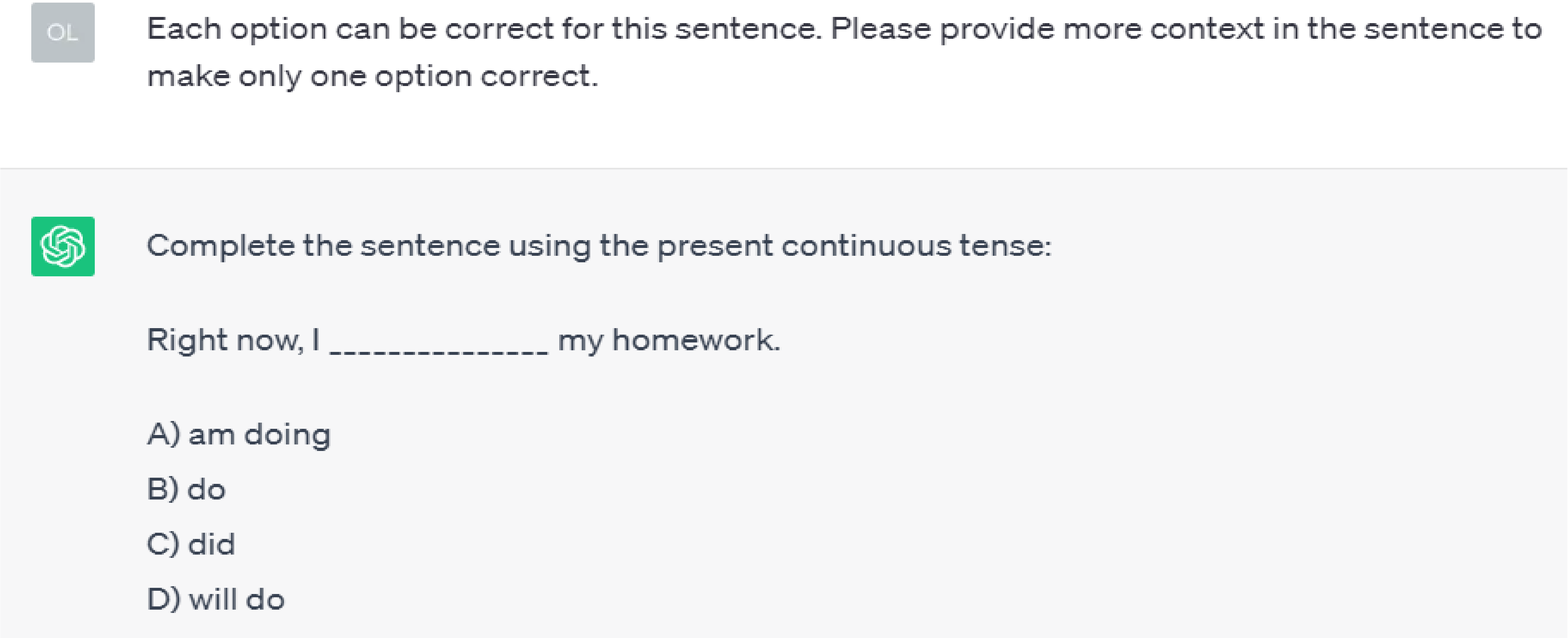
I _______________ my homework.

A) am doing
B) do
C) did
D) will do

ChatGPT, accessed on 25th May 2023

Using ChatGPT for test item generation: Grammar MCQs

ITEM WRITING
for language testing



Each option can be correct for this sentence. Please provide more context in the sentence to make only one option correct.

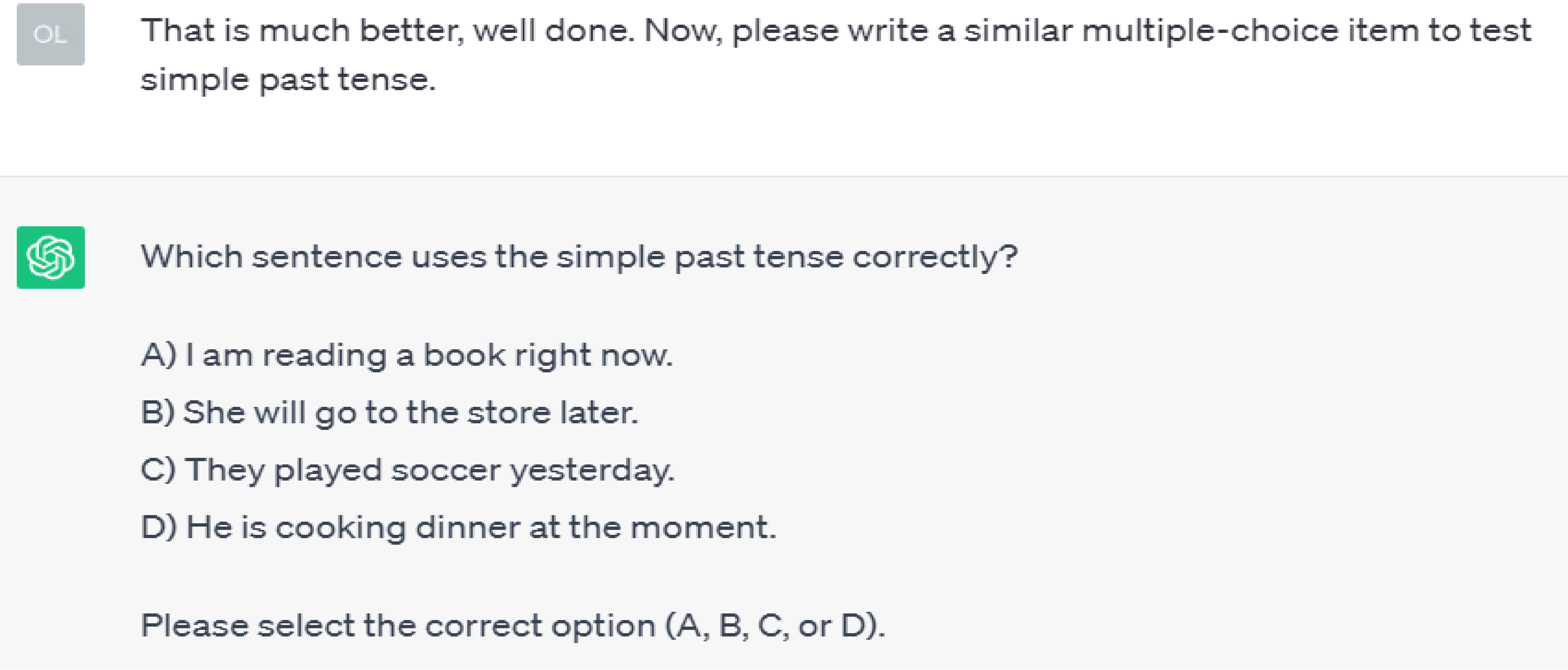
Complete the sentence using the present continuous tense:

Right now, I _______________ my homework.

A) am doing
B) do
C) did
D) will do

ChatGPT, accessed on 25th May 2023

Using ChatGPT for test item generation: Grammar MCQs



That is much better, well done. Now, please write a similar multiple-choice item to test simple past tense.

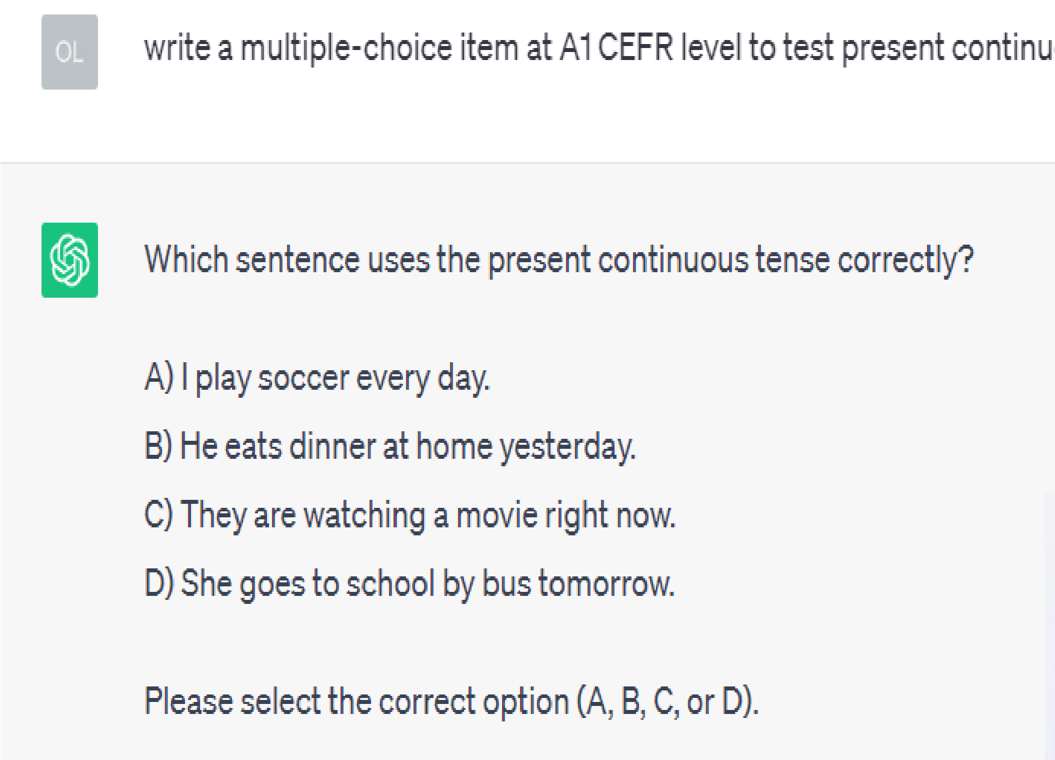
Which sentence uses the simple past tense correctly?

A) I am reading a book right now.
B) She will go to the store later.
C) They played soccer yesterday.
D) He is cooking dinner at the moment.

Please select the correct option (A, B, C, or D).

ChatGPT, accessed on 25th May 2023

Attempt 1



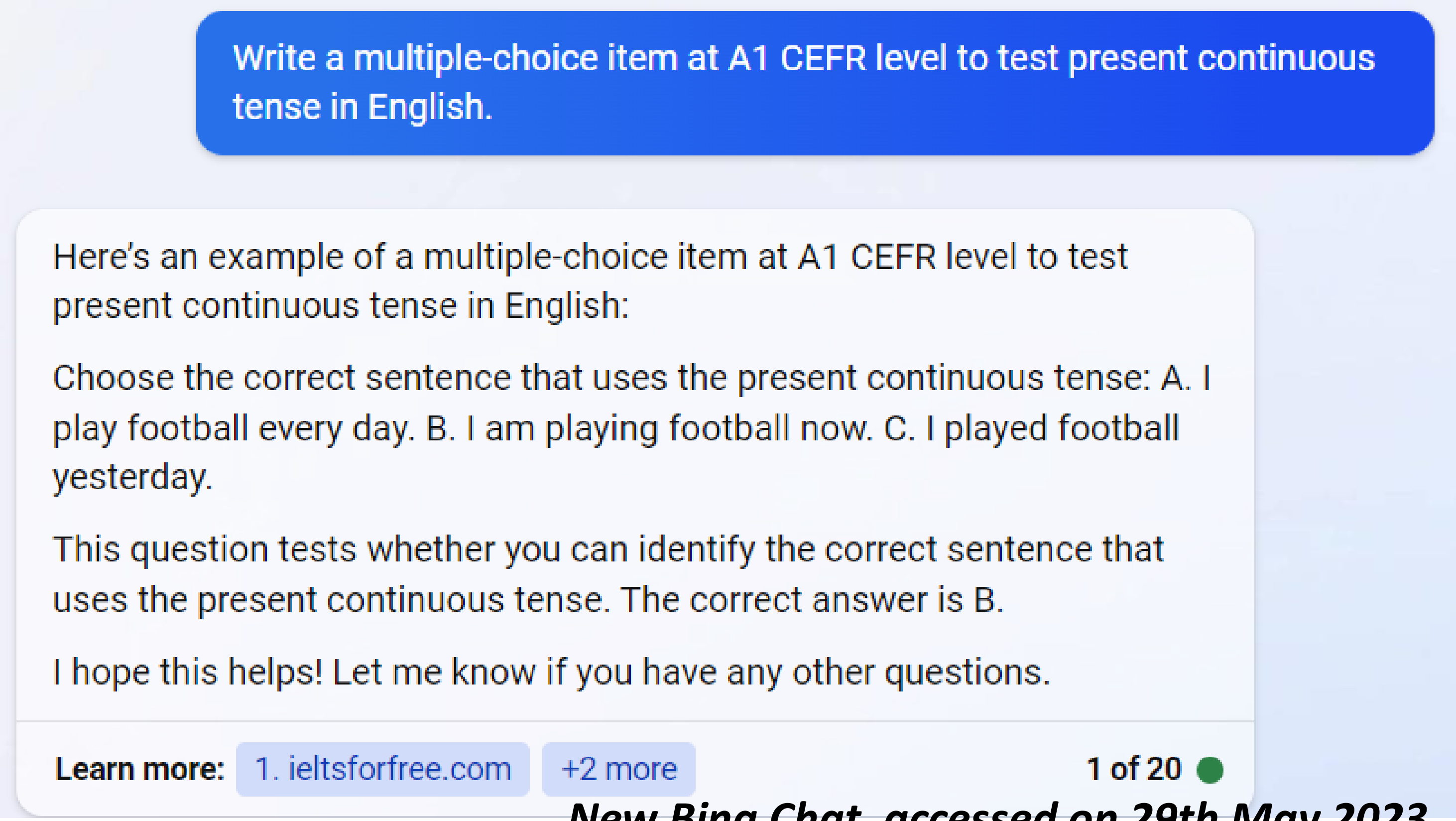
OL write a multiple-choice item at A1 CEFR level to test present continuous tense in English.

Which sentence uses the present continuous tense correctly?

A) I play soccer every day.
B) He eats dinner at home yesterday.
C) They are watching a movie right now.
D) She goes to school by bus tomorrow.

Please select the correct option (A, B, C, or D).

ChatGPT, accessed on 25th May 2023



Write a multiple-choice item at A1 CEFR level to test present continuous tense in English.

Here's an example of a multiple-choice item at A1 CEFR level to test present continuous tense in English:

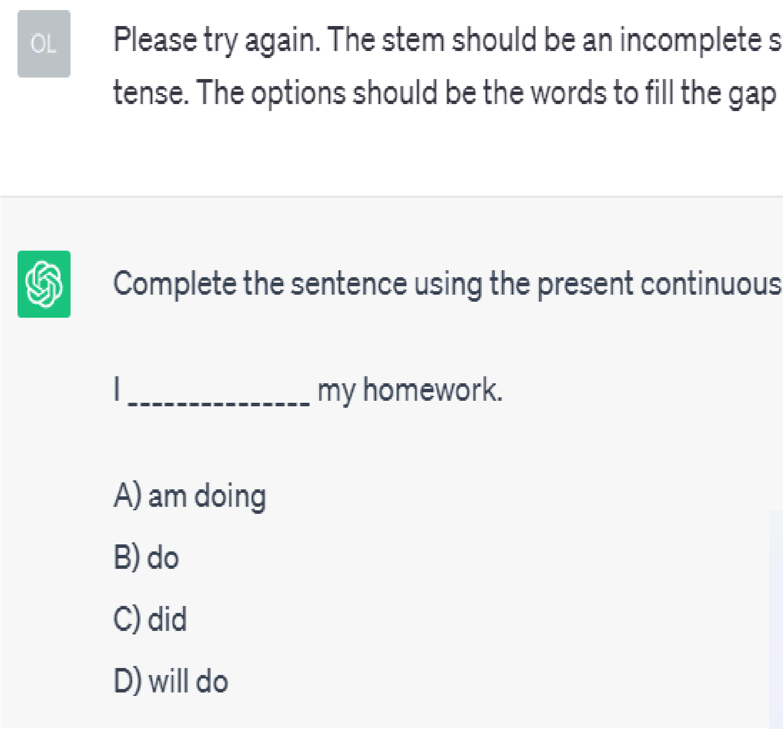
Choose the correct sentence that uses the present continuous tense: A. I play football every day. B. I am playing football now. C. I played football yesterday.

This question tests whether you can identify the correct sentence that uses the present continuous tense. The correct answer is B.

I hope this helps! Let me know if you have any other questions.

New Bing Chat, accessed on 29th May 2023

Attempt 2



Please try again. The stem should be an incomplete sentence that uses present continuous tense. The options should be the words to fill the gap in the sentence.

Complete the sentence using the present continuous tense:

I _____________ my homework.

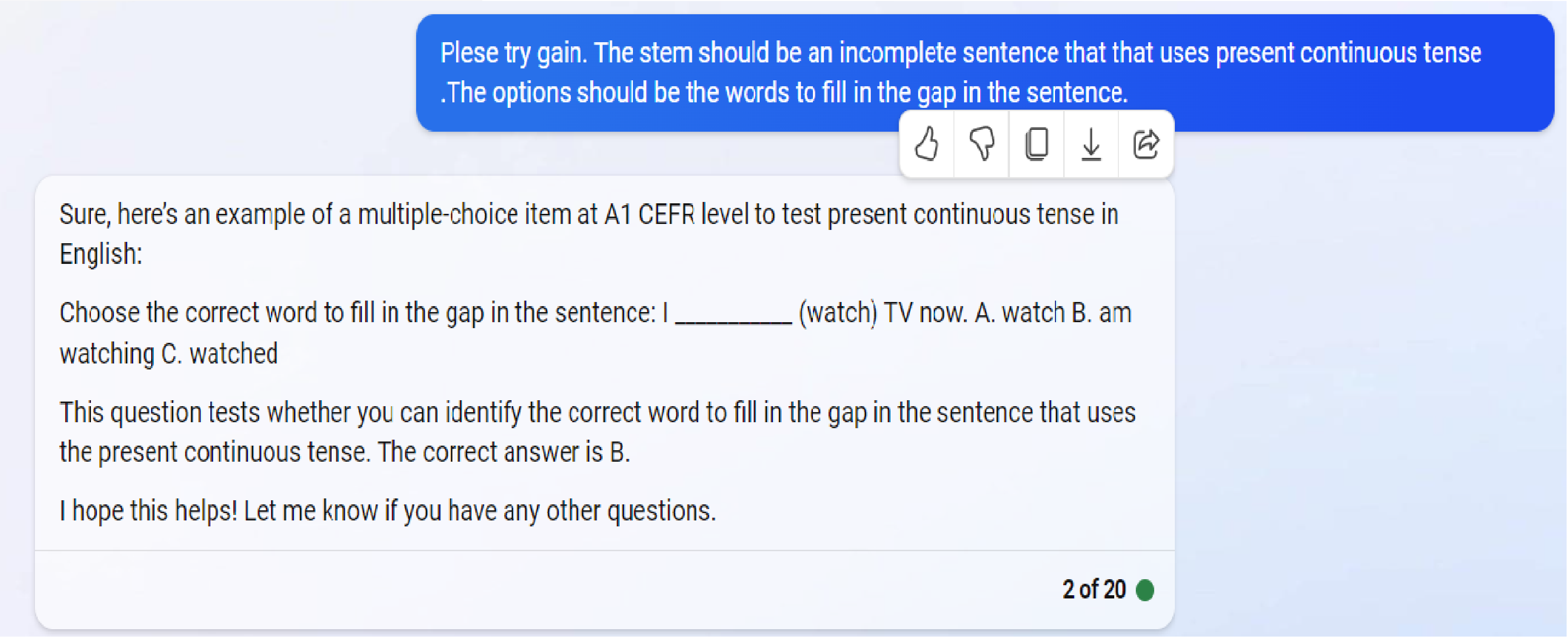
A) am doing

B) do

C) did

D) will do

ChatGPT, accessed on 25th May 2023



Plese try gain. The stem should be an incomplete sentence that that uses present continuous tense .The options should be the words to fill in the gap in the sentence.

Sure, here's an example of a multiple-choice item at A1 CEFR level to test present continuous tense in English:

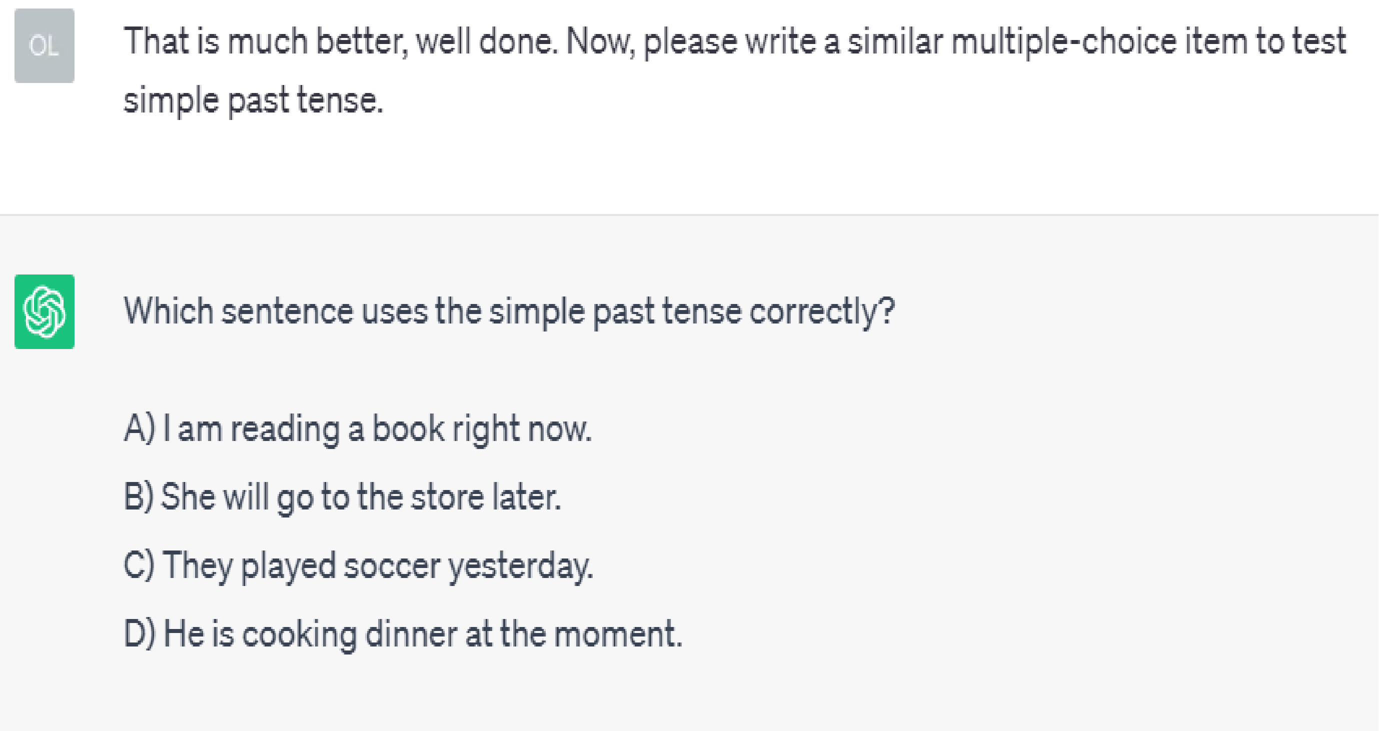
Choose the correct word to fill in the gap in the sentence: I _________ (watch) TV now. A. watch B. am watching C. watched

This question tests whether you can identify the correct word to fill in the gap in the sentence that uses the present continuous tense. The correct answer is B.

I hope this helps! Let me know if you have any other questions.

New Bing Chat, accessed on 29th May 2023

Attempt 3



That is much better, well done. Now, please write a similar multiple-choice item to test simple past tense.

Which sentence uses the simple past tense correctly?

A) I am reading a book right now.

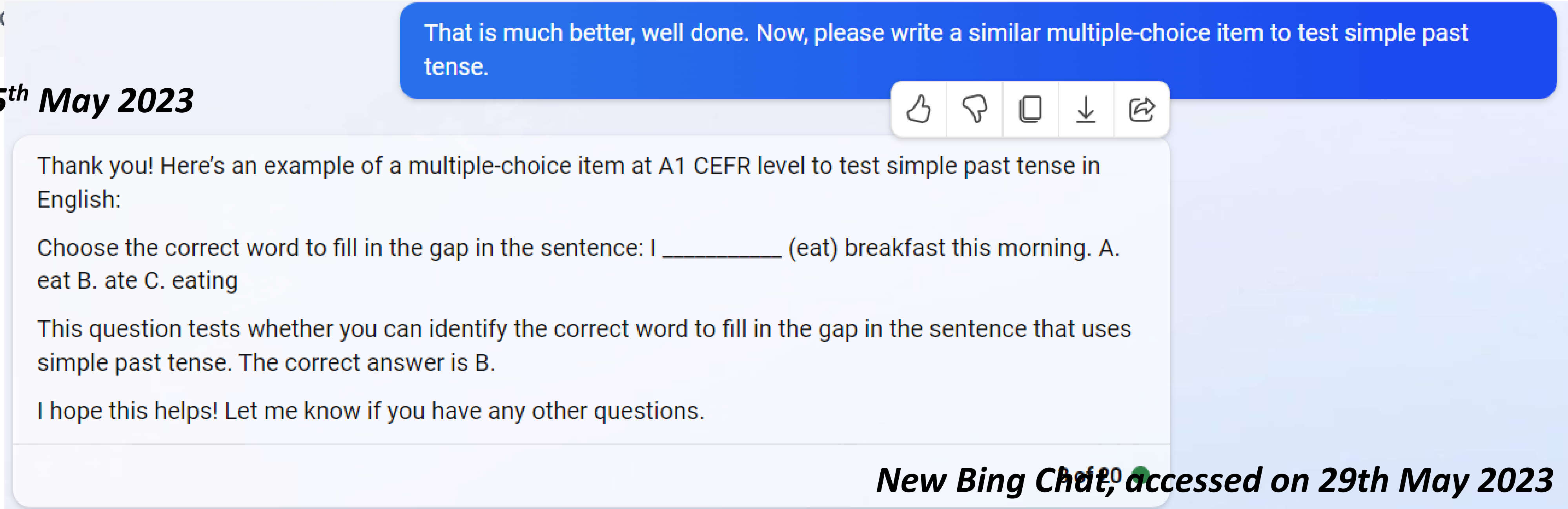
B) She will go to the store later.

C) They played soccer yesterday.

D) He is cooking dinner at the moment.

Please select the correct opti

ChatGPT, accessed on 25th May 2023



That is much better, well done. Now, please write a similar multiple-choice item to test simple past tense.

Thank you! Here's an example of a multiple-choice item at A1 CEFR level to test simple past tense in English:

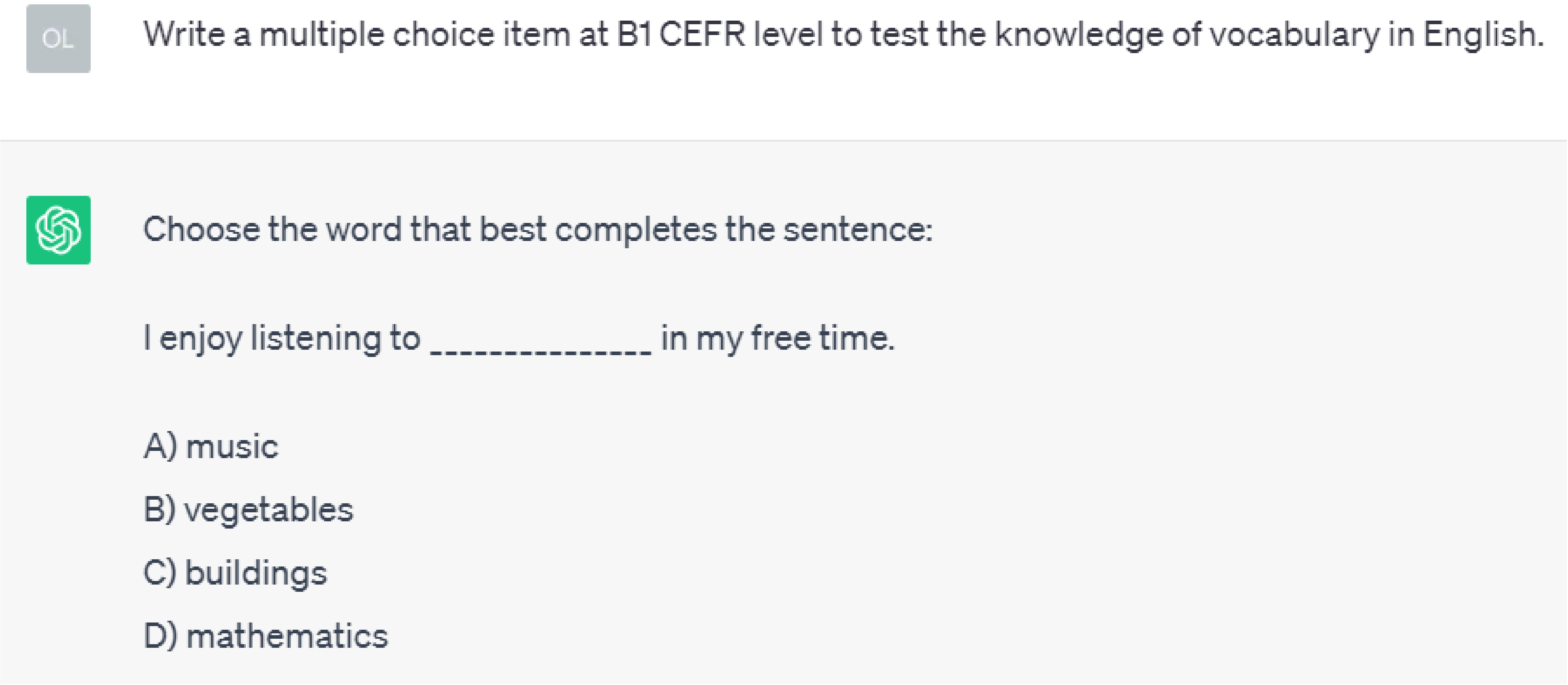
Choose the correct word to fill in the gap in the sentence: I ___________ (eat) breakfast this morning. A. eat B. ate C. eating

This question tests whether you can identify the correct word to fill in the gap in the sentence that uses simple past tense. The correct answer is B.

I hope this helps! Let me know if you have any other questions.

New Bing Chat, accessed on 29th May 2023

Using ChatGPT for test item generation: Vocabulary MCQs



Write a multiple choice item at B1 CEFR level to test the knowledge of vocabulary in English.

Choose the word that best completes the sentence:

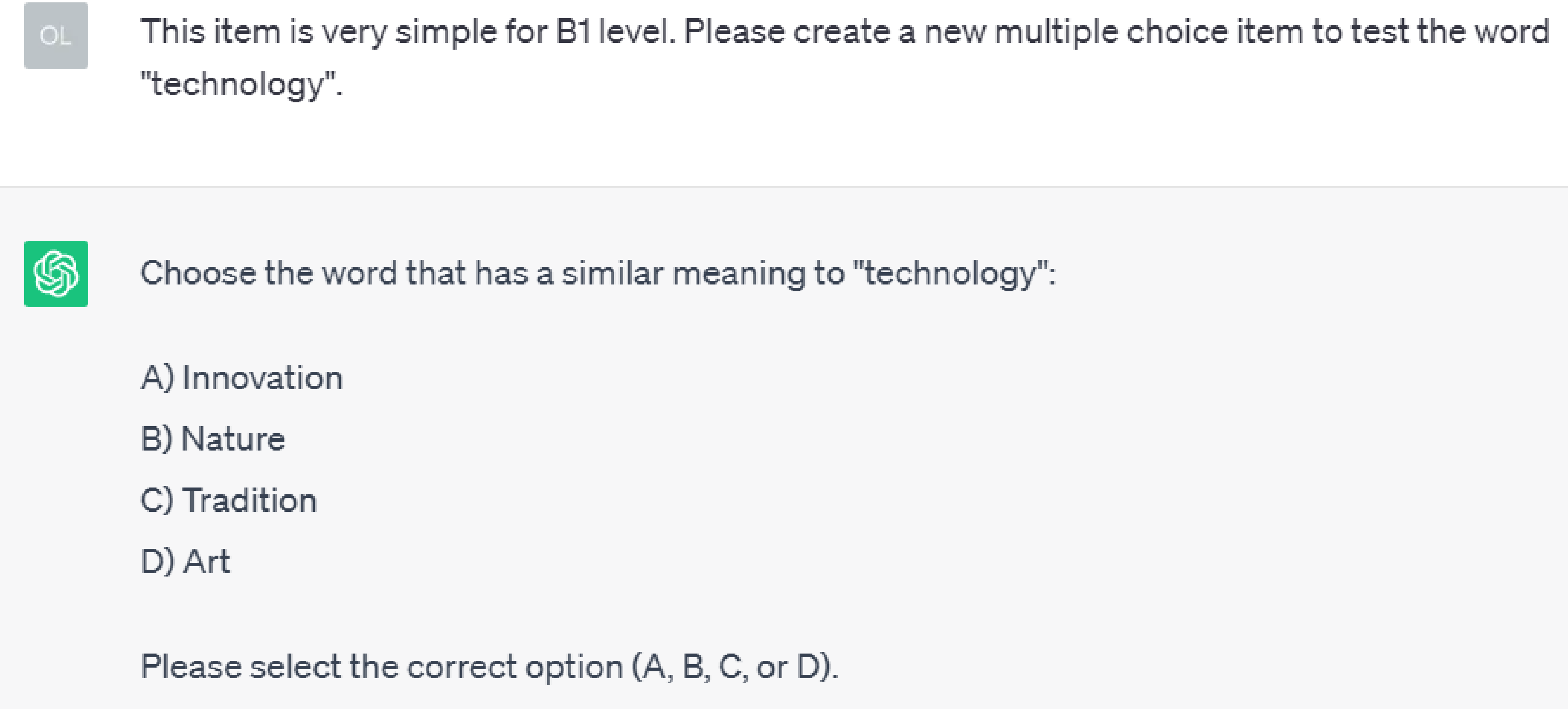
I enjoy listening to _______________ in my free time.

A) music
B) vegetables
C) buildings
D) mathematics

ChatGPT, accessed on 25th May 2023

Using ChatGPT for test item generation: Vocabulary MCQs

ITEM WRITING for language testing



This item is very simple for B1 level. Please create a new multiple choice item to test the word "technology".

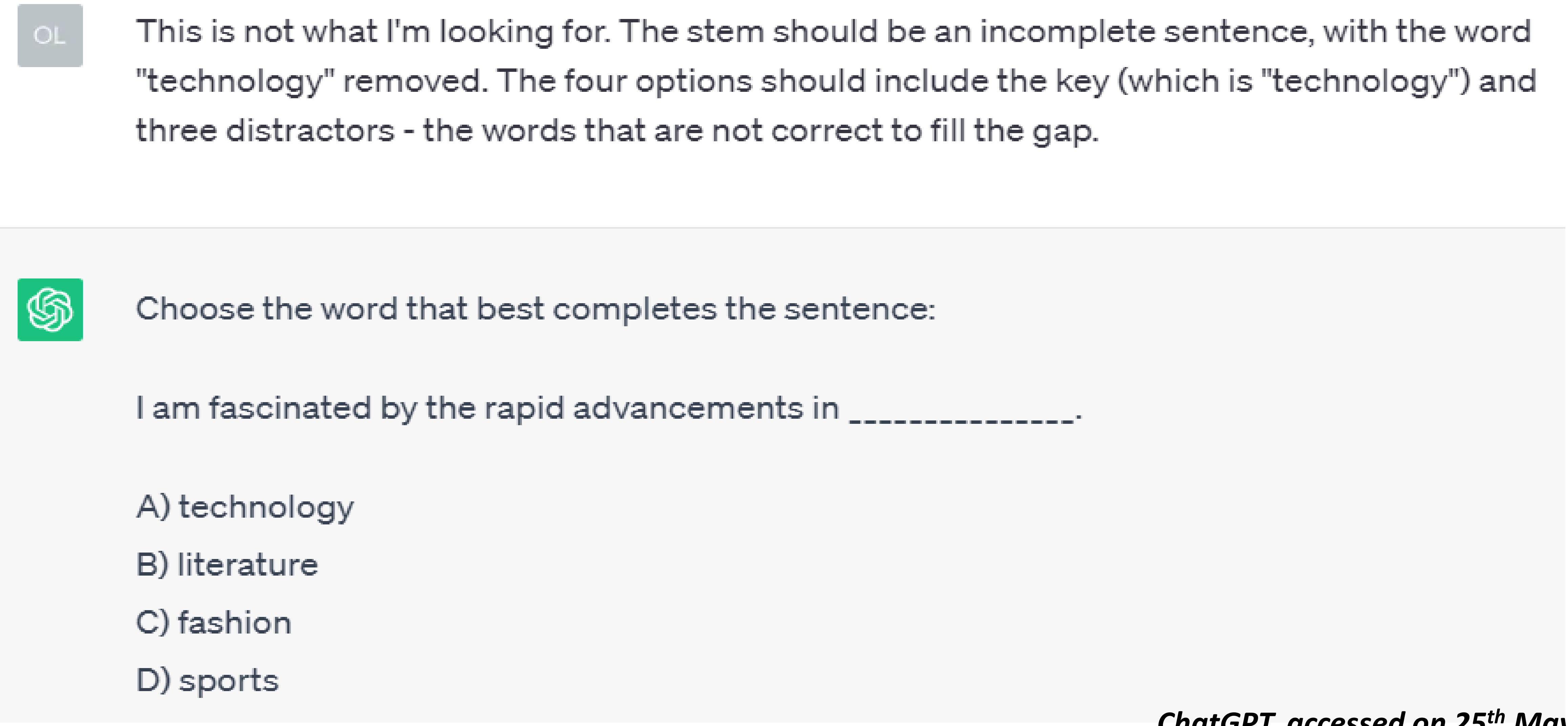
Choose the word that has a similar meaning to "technology":

A) Innovation
B) Nature
C) Tradition
D) Art

Please select the correct option (A, B, C, or D).

ChatGPT, accessed on 25th May 2023

Using ChatGPT for test item generation: Vocabulary MCQs



OL

This is not what I'm looking for. The stem should be an incomplete sentence, with the word "technology" removed. The four options should include the key (which is "technology") and three distractors - the words that are not correct to fill the gap.

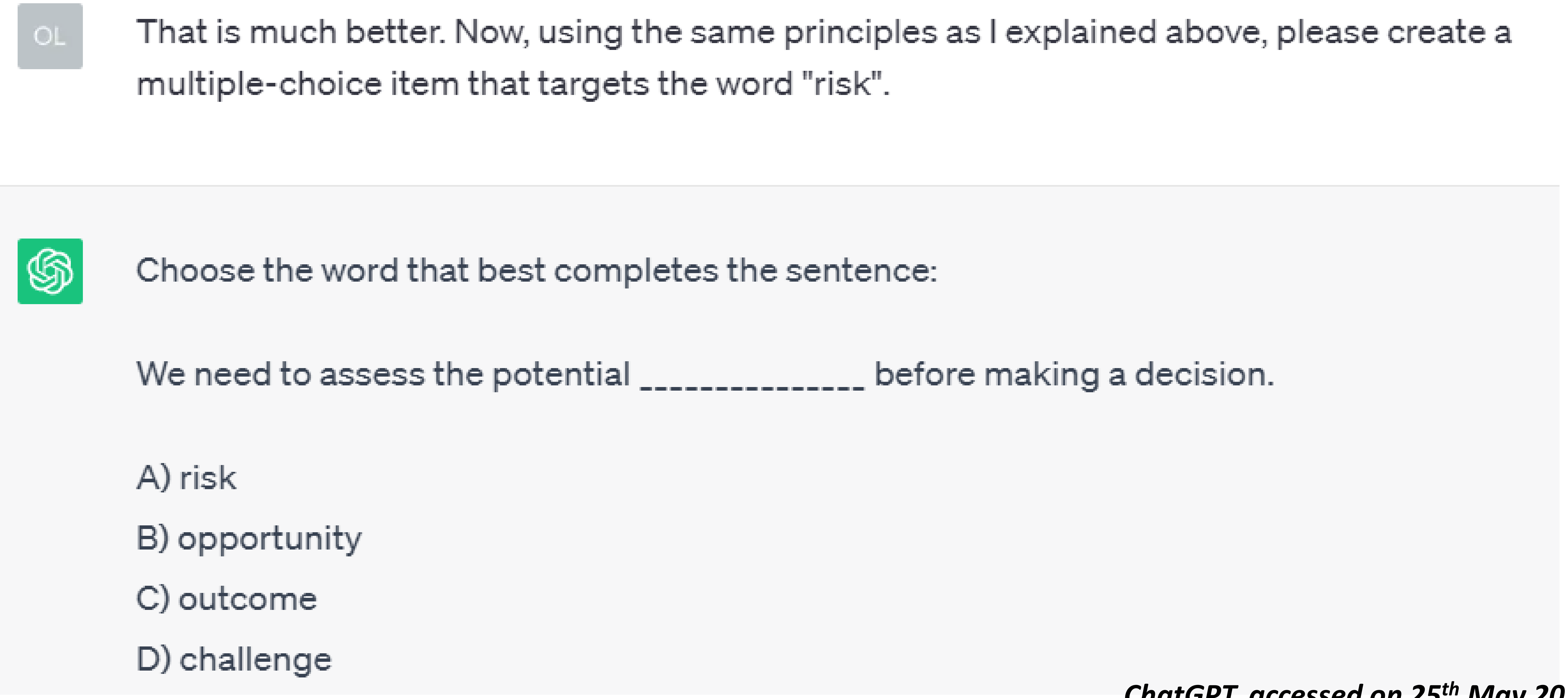
Choose the word that best completes the sentence:

I am fascinated by the rapid advancements in _______________.

A) technology
B) literature
C) fashion
D) sports

ChatGPT, accessed on 25th May 2023

Using ChatGPT for test item generation: Vocabulary MCQs



That is much better. Now, using the same principles as I explained above, please create a multiple-choice item that targets the word "risk".

Choose the word that best completes the sentence:

We need to assess the potential _______________ before making a decision.

A) risk
B) opportunity
C) outcome
D) challenge

ChatGPT, accessed on 25th May 2023

Using ChatGPT for test item generation: Reading

Prompt: Generate an 800 word text on a general academic topic and six reading comprehension questions at the B2 level of the Common European Framework of Reference

Topic: The Impact of Technology on Education

In recent years, technology has had a significant impact on the education sector. With the emergence of digital devices such as laptops, smartphones, and tablets, education has become more accessible, personalized, and interactive. In this article, we will explore how technology has influenced the way we learn and discuss the benefits and challenges associated with its use in the classroom.

One of the most significant impacts of technology on education is that it has made learning more accessible. Students can now access learning resources and educational materials from anywhere, at any time, using digital devices. *Etc.*

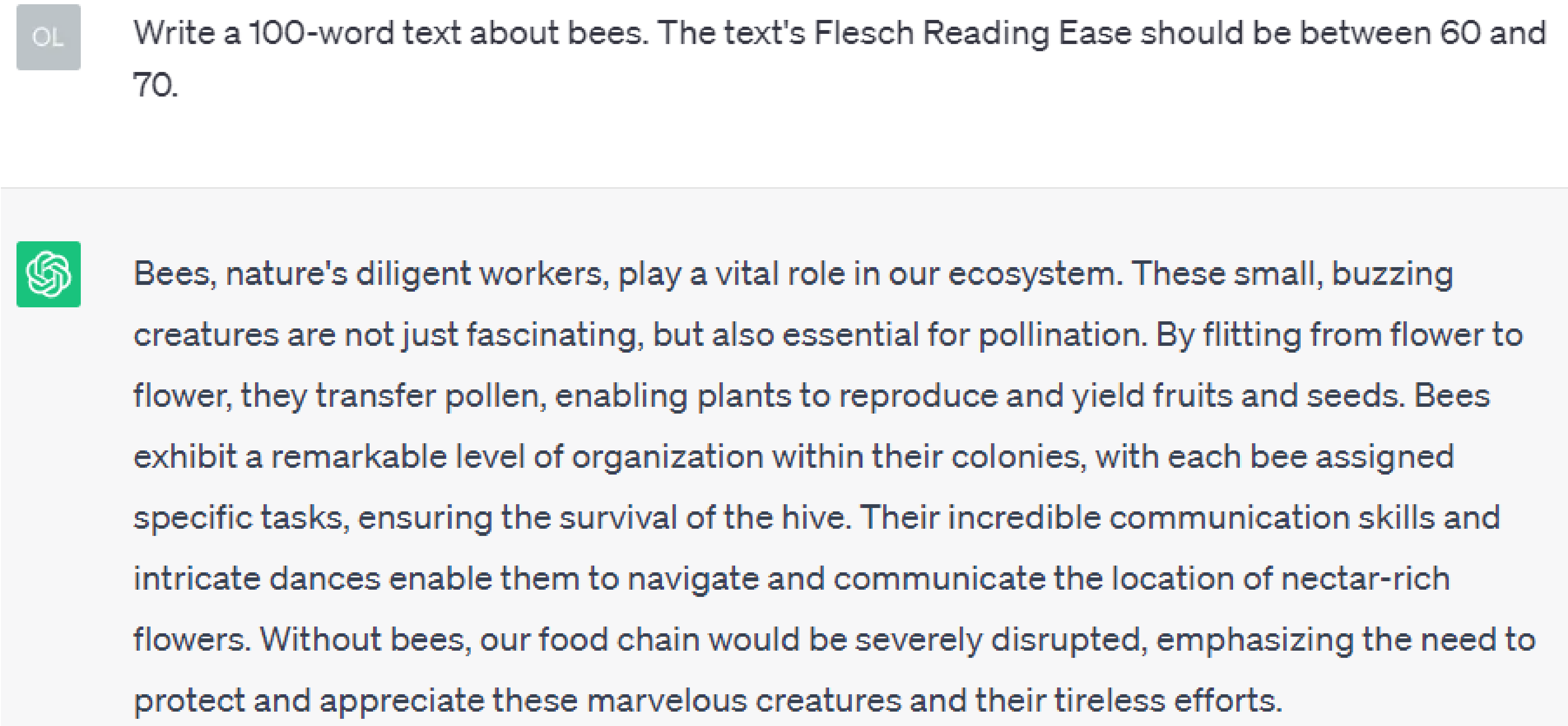
1. What impact has technology had on education?
    A. No impact at all
    B. A significant impact
    C. A minor impact
    D. An unknown impact

2. How has technology made learning more accessible?
    A. It has made learning less accessible
    B. It has provided access to educational resources from anywhere and anytime
    C. It has made learning more expensive
    D. It has made learning more difficult to access

Etc.

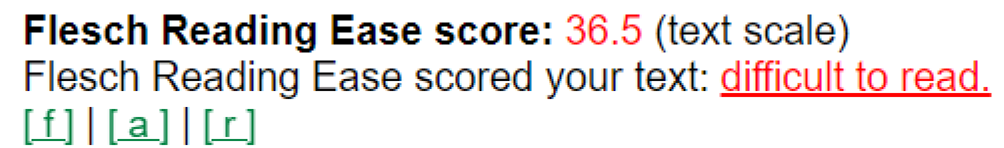
Brunfaut, 2023

Using ChatGPT for test item generation: Reading



OL Write a 100-word text about bees. The text's Flesch Reading Ease should be between 60 and 70.

Bees, nature's diligent workers, play a vital role in our ecosystem. These small, buzzing creatures are not just fascinating, but also essential for pollination. By flitting from flower to flower, they transfer pollen, enabling plants to reproduce and yield fruits and seeds. Bees exhibit a remarkable level of organization within their colonies, with each bee assigned specific tasks, ensuring the survival of the hive. Their incredible communication skills and intricate dances enable them to navigate and communicate the location of nectar-rich flowers. Without bees, our food chain would be severely disrupted, emphasizing the need to protect and appreciate these marvelous creatures and their tireless efforts.



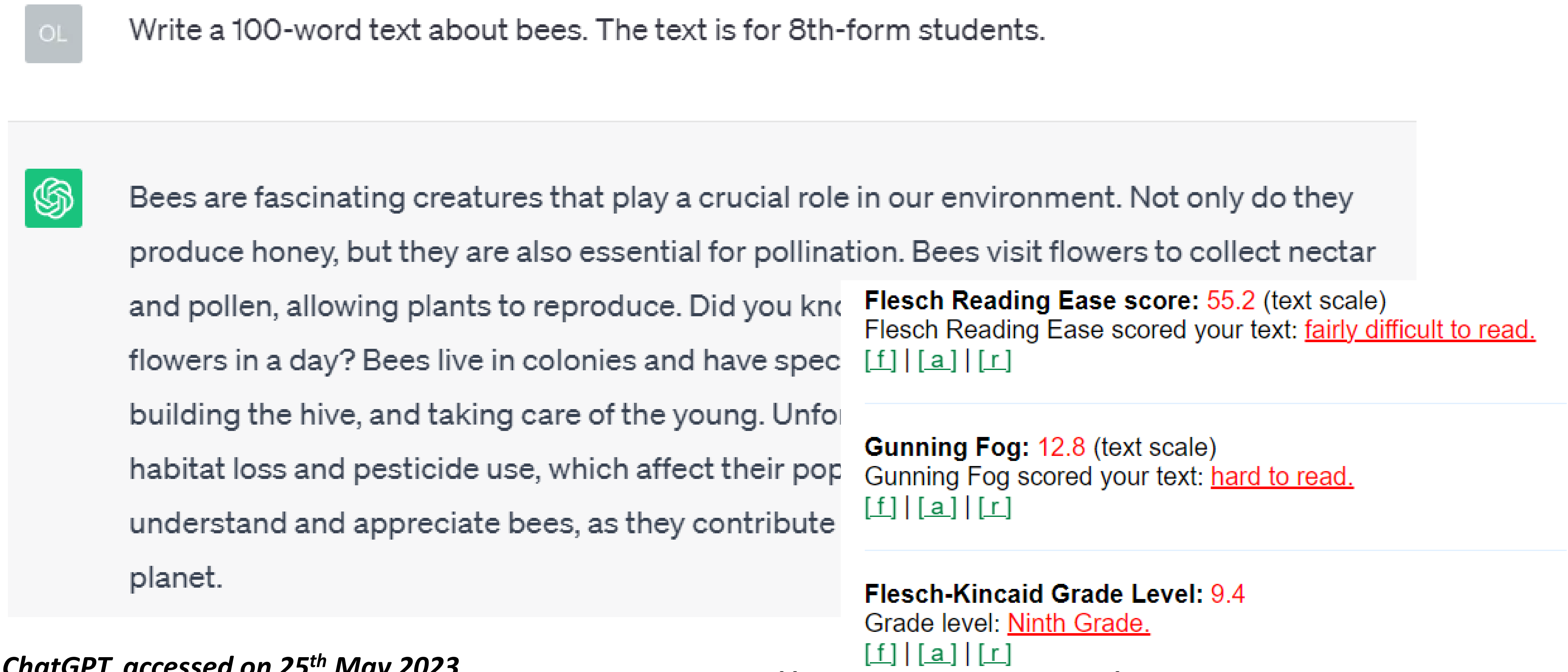
Flesch Reading Ease score: 36.5 (text scale)
Flesch Reading Ease scored your text: difficult to read.
[f] | [a] | [r]

ChatGPT, accessed on 25th May 2023

https://readabilityformulas.com/free-readability-formula-tests.php

Using ChatGPT for test item generation: Reading

ITEM WRITING
for language testing



OL

Write a 100-word text about bees. The text is for 8th-form students.

Bees are fascinating creatures that play a crucial role in our environment. Not only do they produce honey, but they are also essential for pollination. Bees visit flowers to collect nectar and pollen, allowing plants to reproduce. Did you kn
flowers in a day? Bees live in colonies and have spec
building the hive, and taking care of the young. Unfo
habitat loss and pesticide use, which affect their pop
understand and appreciate bees, as they contribute
planet.

Flesch Reading Ease score: 55.2 (text scale)
Flesch Reading Ease scored your text: fairly difficult to read.
[f] | [a] | [r]

Gunning Fog: 12.8 (text scale)
Gunning Fog scored your text: hard to read.
[f] | [a] | [r]

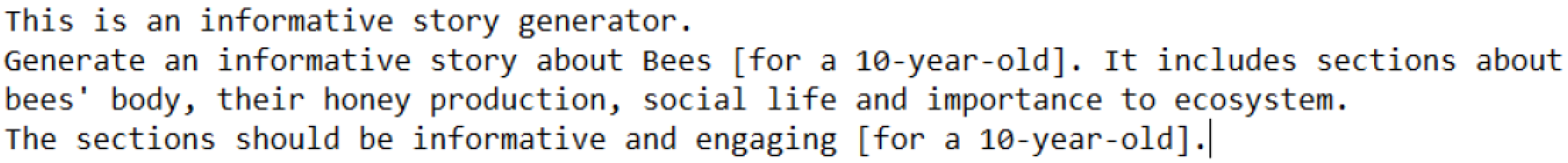
Flesch-Kincaid Grade Level: 9.4
Grade level: Ninth Grade.
[f] | [a] | [r]

ChatGPT, accessed on 25th May 2023

https://readabilityformulas.com/free-readability-formula-tests.php

Bezirhan & von Davier (2023)

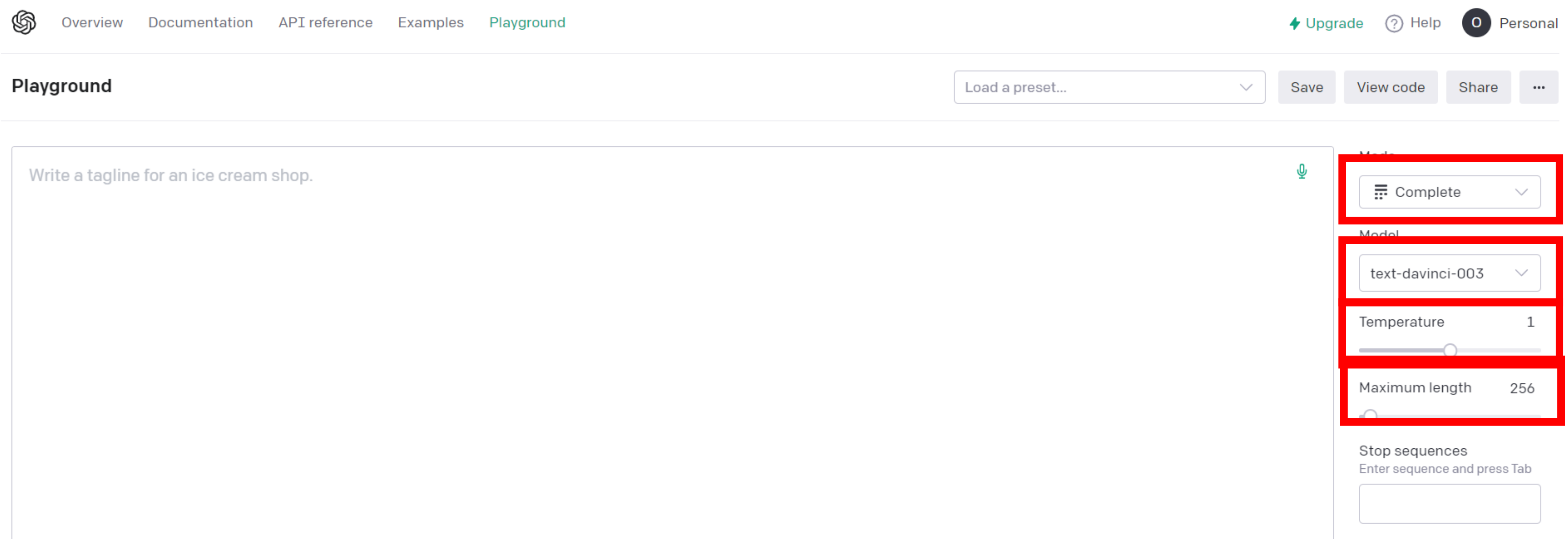
- texts for the primary school reading achievement test PIRLS
- “*prompt design is essential for …GPT… because it directly affect the quality and relevance of the output*” (Bezirhan & von Davier, 2023, p.15)
- “zero-shot” and “one-shot” learning modes



```
This is an informative story generator.
Generate an informative story about Bees [for a 10-year-old]. It includes sections about
bees' body, their honey production, social life and importance to ecosystem.
The sections should be informative and engaging [for a 10-year-old].
```

(Bezirhan & von Davier, 2023, p.28)

- Python to send API requests with prompt design and parameters to the OpenAI's API service
- Open AI's API Playground https://platform.openai.com/playground try for free ($18 worth tokens)



Write a tagline for an ice cream shop.

Mode: Complete

Model: text-davinci-003

Temperature 1

Maximum length 256

Stop sequences
Enter sequence and press Tab

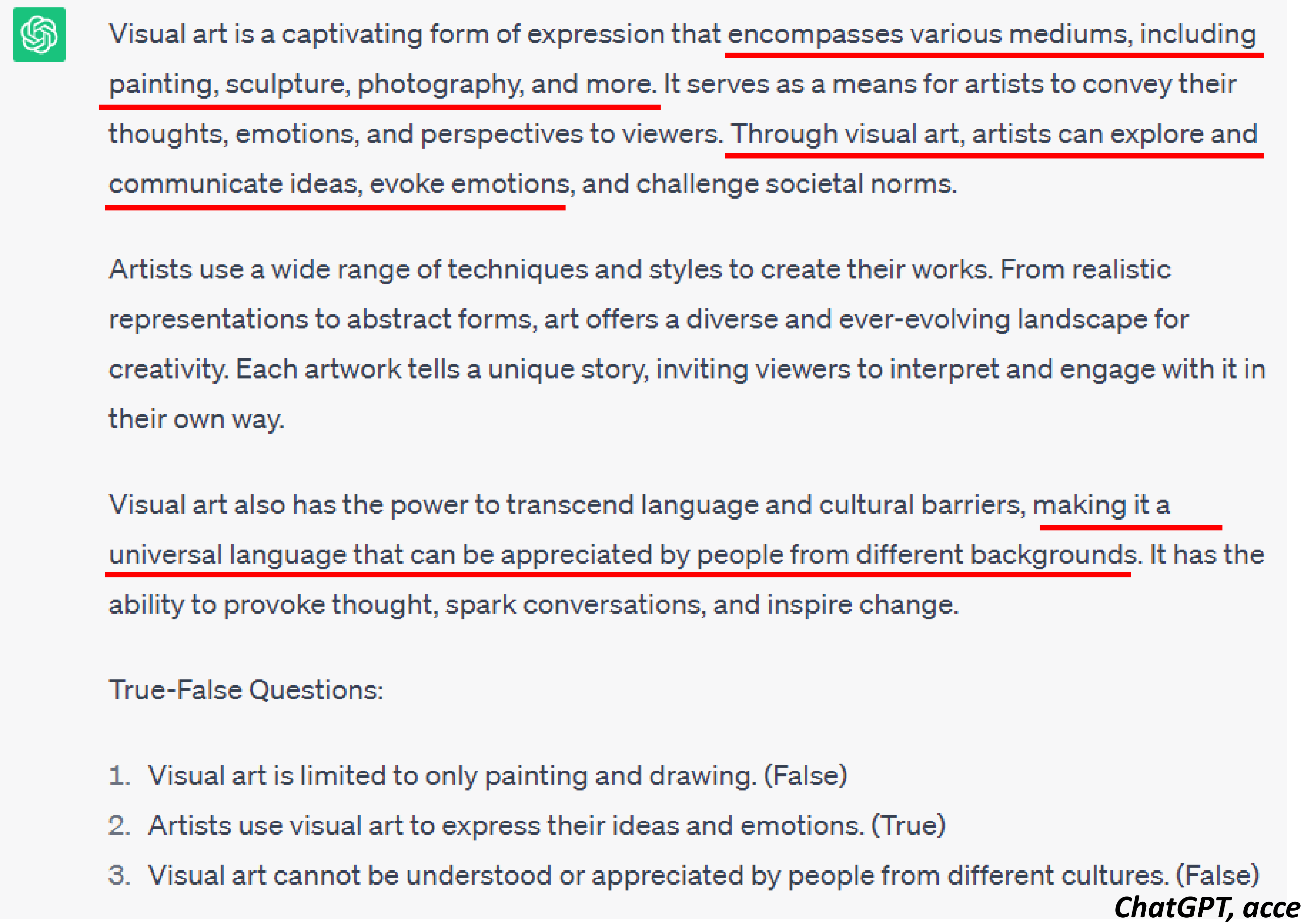
Open AI, API Playground, accessed on 28th May 2023

Bezirhan & von Davier (2023)

“inclusion of grade/age information reduced the text difficulty score for the generated passages” (Bezirhan & von Davier, 2023, p.19)

- Blinded expert judgement of passage quality

	Informative passages		**Fiction stories**	
	original	GPT-generated	original	GPT-generated
adequate	96%	92%	80%	96%
engaging	86%	84%	74%	94%
coherent	94%	84%	88%	88%



Visual art is a captivating form of expression that encompasses various mediums, including painting, sculpture, photography, and more. It serves as a means for artists to convey their thoughts, emotions, and perspectives to viewers. Through visual art, artists can explore and communicate ideas, evoke emotions, and challenge societal norms.

Artists use a wide range of techniques and styles to create their works. From realistic representations to abstract forms, art offers a diverse and ever-evolving landscape for creativity. Each artwork tells a unique story, inviting viewers to interpret and engage with it in their own way.

Visual art also has the power to transcend language and cultural barriers, making it a universal language that can be appreciated by people from different backgrounds. It has the ability to provoke thought, spark conversations, and inspire change.

True-False Questions:

1. Visual art is limited to only painting and drawing. (False)
2. Artists use visual art to express their ideas and emotions. (True)
3. Visual art cannot be understood or appreciated by people from different cultures. (False)

ChatGPT, accessed on 25th May 2023

Using GPT for test item generation: Duolingo

Interactive reading task

Based on a single short text to save testing time

The text is generated by GPT

1. Vocabulary in context: multiple-choice gap fill
2. Text completion: filling in a missing sentence, multiple choice
3. Comprehension questions: open-ended, text highlighting
4. Main idea: multiple-choice
5. Title: multiple-choice

Candidate words for deletion are computed by a language model [BERT?] and "*filtered based on the likelihood*" (Attali et al, 2022, p.5)

"*Ideally, successful distractors have low, but not too low, likelihood, and have the same syntactic part-of-speech as the correct answer.*" (Attali et al, 2022, p.5)

Attali et al., 2022

Using GPT for test item generation: Duolingo

Interactive reading task

Based on a single short text to save testing time

The text is generated by GPT

1. Vocabulary in context: multiple-choice gap fill
2. Text completion: filling in a missing sentence, multiple choice
3. Comprehension questions: open-ended, text highlighting
4. Main idea: multiple-choice
5. Title: multiple-choice

GPT is not good at generating distractors for multiple-choice items (Attali et al, 2023)

NLP metrics: average similarity to the correct answer, to the source passage, and to individual sentences in the source passage (Attali et al, 2022)

Attali et al., 2022

Using GPT for test item generation: Duolingo

Interactive reading task

Based on a single short text to save testing time

The text is generated by GPT

1. Vocabulary in context: multiple-choice gap fill
2. Text completion: filling in a missing sentence, multiple choice
3. Comprehension questions: open-ended, text highlighting
4. Main idea: multiple-choice
5. Title: multiple-choice

Questions generated by GPT.

“To better guarantee that the generated questions are answerable using the passage, we use an external question answering model to predict the overall likelihood that the question can be answered” (Attali et al, 2022, p.5)

Attali et al., 2022

Using GPT for test item generation: Duolingo

Interactive reading task

Based on a single short text to save testing time

The text is generated by GPT

1. Vocabulary in context: multiple-choice gap fill
2. Text completion: filling in a missing sentence, multiple choice
3. Comprehension questions: open-ended, text highlighting
4. Main idea: multiple-choice
5. Title: multiple-choice

Multiple potential answers are generated by GPT and evaluated automatically "*based on their similarity to the passage and average negative log likelihood as estimated by the language model*" (Attali et al, 2022)

Attali et al., 2022

Using large language models for test item generation

Q1: How efficiently can we generate test items with GPT?

- Large-scale IR task pilot: 14,000 → 789 (5.6%) → 454 (3.2%) (Attali et al, 2022)

Q2: How good is the quality of the test items generated with GPT?

- Many items are of low quality (judging by 96.8% rejection level)
- Are the remaining 3.2% good enough to be used for testing? E.g., can the model generate strong and plausible distractors based on the likelihood statistics?

Q3: How much human review is needed for the GPT-generated items?

- *"The review process is estimated to have taken about 15 min per passage (including questions), across all rounds of review"* (Attali et al, 2022, p.7)

Using large language models for large-scale operational item writing

- What is possible to be done with AI?
- What is not possible, because AI is not good at producing items of this type?
- Can we replace this particular item type with a different one that lands itself better to automated generation?
- If we do so, will this affect the construct of testing?
- What will the quality of the resulting items be, if we use AI?
- How much review time/iterations do we need to build into the system? (fact checking, bias review, editing)

Using large language models for classroom-based testing

ITEM WRITING
for language testing

- Generate linguistically accurate, topical, coherent texts
- Generate some (!) items to go with the text (e.g., comprehension questions targeting important detail)
- BUT correct prompting is important
- BUT text difficulty should be checked
- BUT items need editing
- BUT some items (e.g., targeting inference, global understanding, synthesis) might have to be teacher-written

AI in language testing SIG

- Are you a member of EALTA?
- Do you want to learn more about using AI for language testing?
- Do you think “AI in language testing” SIG might be a good idea?
- Are you interested to take an active part in setting up and running the SIG?

Please write to Olena Rossi olena.rossi@itemwriting.co

itemwriting.co

olena.rossi@itemwriting.co

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,* July 2021. https://dl.acm.org/doi/abs/10.1145/3461702.3462624

Attali, Y., LaFlair, G., & Runge, A. (2023, March 31). A new paradigm for test development [Duolingo webinar series]. https://www.youtube.com/watch?v=rRc96oe9bzk&t=2s

Attali, Y., Runge, A., LaFlair, G.T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*, 903077. https://doi.org/10.3389/frai.2022.903077

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with Open AI's large language model. Preprint. https://doi.org/10.48550/arXiv.2304.04616

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020).* https://proceedings.neurips.cc/paper/2020/hash/ 1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Brunfaut, T. (2023). Back to basics to be fit for the digital age: Reflecting on advancements, opportunities and challenges of technology-enhanced language assessment. ALTE 8th International Conference, April 2023, Madrid. https://www.alte.org/Madrid-2023

Christodoulou, D. (2023, February 9). How good is ChatGPT at writing essays? Some data! https://blog.nomoremarking.com/how-good-is-chatgpt-at-writing-essays-some-data-eda60de7aee5

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. Educational Measurement: Issues and Practice, 32, 36-50. https://doi.org/10.1111/emip.12018

Li, L., & Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. Proceedings of the Third Workshop on Narrative Understanding, NUSE 2021. https://aclanthology.org/2021.nuse-1.5

References

Logé, C., Ross, E., Dadey, D. Y. A., Jain, S., Saporta, A., Ng, A. Y., & Rajpurkar, P. (2021). Q-pain: A question answering dataset to measure social bias in pain management. Preprint. https://doi.org/10.48550/arXiv.2108.01764

Shihadeh, J., Ackerman, M., Troske, A., Lawson, N., & Gonzalez, E. (2022). Brilliance bias in GPT-3. IEEE Global Humanitarian Technology Conference, September 2022. https://doi.org/10.1109/GHTC55712.2022.9910995

Yoon, S.-Y., Lee, C.M., Houghton, P., Lopez, M., Sakano, J., Loukina, A., Krovetz, B., Lu, C., & Madnani, N. (2017). Analyzing item generation with natural language processing tools for the TOEIC® Listening Test. ETS Research Report Series, 2017, pp. 1-9. https://doi.org/10.1002/ets2.12183

Zong, M., & Krishnamachari, B. (2022). A survey on GPT-3. Preprint. https://doi.org/10.48550/arXiv.2212.00857

itemwriting.co

THANK YOU!

olena.rossi@itemwriting.co