

Evaluating the effectiveness of Surpass Copilot: an AI-assisted distractor generator

1. Objectives.....	1
2. Overview of the trial.....	2
3. Key findings.....	2
4. Recommendations.....	4
5. Proposed next steps	6
Appendix 1.....	7
Appendix 2.....	9
Appendix 3.....	25

1. Objectives

Surpass Co-pilot is an AI-powered tool integrated within the Surpass Platform, developed to generate distractor options for multiple choice questions (MCQ). It is designed to enhance and accelerate the process of item writing by supporting Subject Matter Experts (SMEs) in writing high-quality multiple-choice questions.

Surpass Copilot key functionality includes:

- Generating distractors and providing rationale for each distractor.
- Enabling SMEs to input specific instructions or prompts to generate more effective distractors.

SQA has access to this tool through SOLAR, SQA's online assessment system.

The purpose of the trial was to evaluate the effectiveness and accuracy of the AI-generated distractors compared to traditional, human-written test items. It aimed to explore the following key aspects:

Quality of distractors

- Are generated distractors plausible and relevant to the questions?

Item difficulty

- Do the distractors match the intended difficulty level of the questions?

Time efficiency

- Would SMEs need to make significant adjustments to make distractors fit for purpose?
- Would Surpass Copilot enhance the item development process and reduce human resource workload?

The following success criteria scale was used to indicate the performance of the tool:

Scale		Description
1	Poor	Distractors fail to meet expectations, there are significant issues, requiring extensive human intervention
2	Fair	The AI shows some benefits, but there are notable deficiencies in efficiency, quality and accuracy, leading to a less reliable output
3	Good	Distractors meet the basic requirements with a few minor issues. The quality is generally satisfactory with some areas requiring improvement
4	Very good	Distractors are of very good quality, with minor issues that would only require some human intervention. There are clear benefits to the use of AI.
5	Excellent	Distractors are of high quality, with minimal or no issues. The AI exceeds expectations.

More detailed information on the criteria used can be found in the *SME feedback form* (See Appendix 1)

2. Overview of the trial

The trial focused on evaluating the effectiveness of the AI tool for two subjects:

- Higher Physics
- ESOL: National 5 ESOL Reading and Skills for English B1 Reading test

Two internal SMEs were involved in the trial. The SMEs were responsible for:

- reviewing the AI-generated distractors to determine their plausibility and effectiveness
- developing prompts to steer the development of AI-generated distractors

We used past papers in the trial to access pre-existing, quality-assured questions and distractors.

Key steps in the trial:

1. Uploading of questions
 - Questions and correct answers from past papers were uploaded into the system
2. Distractor generation:
 - Two versions of tests were created: one without prompts and one with SME-provided prompts
3. Review and Analysis of distractors
 - SMEs reviewed generated distractors, analysing their quality and relevance

3. Key findings

SMEs chose ratings of 1 for Higher Physics and 2 for ESOL:

1	Poor	Distractors fail to meet expectations, there are significant issues, requiring extensive human intervention
2	Fair	The AI shows some benefits, but there are notable deficiencies in efficiency, quality and accuracy, leading to a less reliable output

In its current iteration, the AI tool still requires substantial input from SMEs to produce the items. The tool offers limited value and would likely increase the time spent on reviewing and rewriting items to meet the required standards.

Below is the list of key insights:

Quality of distractors:

- The overall quality was inconsistent, with many distractors being implausible or irrelevant to the question context.
- While justification may appear correct on the surface level, the underlying logic can sometimes be flawed (both with and without prompts). In Physics, it generates distractors that are not always based on the data provided in the question.
- At times, the tool generates distractors that are actually correct answers, stating in its justification that the option could distract the learner by 'being a less commonly known answer.'
- The tool performed slightly better with traditional question formats (direct questions), but encountered difficulties with sentence completion questions and elliptical statements, often misunderstanding the context and generating random or irrelevant responses (especially if the statement was not followed by ellipses).
- Distractors were not reflective of the expected thought process of the test taker (with and without prompts). The tool fails to replicate how a learner would approach tasks, and as a result, it generates items that are not very convincing.
- Some sources used to justify responses, such as Wikipedia or some blogs, could not be considered reliable.

Item Difficulty and discrimination

- Distractors do not always align with the intended level of difficulty, even when specific guidance is provided in the prompt box. The tool fails to produce distractors that are appropriately discriminating.

Time Efficiency

- Due to the quality of distractors, significant adjustments would be required to make distractors fit for purpose.
- The system doesn't appear to be able to generate responses in scientific notation, resulting in lots of editing being required.

Other

- **Justifications:** The system does not save prompts or rationales for distractors, meaning SMEs would not have access to these at later stages of item development or the quality assurance process. Currently, the only way to access these is by taking screenshots.
- **Test structure:** In its current iteration, the system struggles to manage the typical SQA National 5 ESOL test structure, where 10 questions follow one long text. This is likely a limitation of the test platform rather than the AI tool itself. It handles shorter texts better, producing slightly better-quality distractors. In Physics, there are many formatting issues, such as no access to certain math

symbols. Since the MathType widget cannot be used in the prompts box, it significantly restricts the range of prompts.

- **Visual data:** For subjects requiring graphical content, the tool is currently restrictive as it lacks the capability to generate graphs within the answer options and cannot interpret or incorporate graphs in questions.
- **Prompts:** The current approach of using a single prompt box may lead to confusion and inefficiencies, as prompts related to formatting and those targeting specific domain knowledge are mixed together. This risks unclear or inconsistent outputs and poses a challenge for scalability. As more item writers get involved, it could become harder to ensure a consistent approach across users. Mixing different types of prompts makes the process more prone to errors or oversight.

See *Appendices 2 and 3* for examples of prompts and AI-generated distractors.

4. Recommendations

The following list highlights key areas that would need to be addressed to support item writing for SQA assessments effectively.

Justifications and feedback

- Provide an option to save rationales and suggested distractors for use at later stages of item development and quality assurance, ensuring easy access and reusability.
- Ensure that justifications clearly and directly address the instructions in the prompt box, improving the alignment between input and output.
- Enhance the feedback mechanism by allowing users to flag and annotate issues with distractors and enabling the tool to learn and improve based on SMEs' input. (Copilot 2.0 will allow sharing feedback however it won't be interactive and will be limited to reporting items that are incorrect, inappropriate or offensive.)
- Consider implementing adaptive difficulty scaling, so the tool can adjust the complexity of the questions based on user feedback and requirements.

Prompts

- It might be beneficial to introduce two separate boxes for prompts:
 - Formatting prompts, which could be standardised and reusable across different questions and tests, ensuring consistency and reducing the need for repetitive input.
 - Domain-specific prompts, which would focus on the subject matter and difficulty level.
- Incorporate a functionality to create and maintain a bank of prompts, including generic and subject-specific prompts, for efficient, consistent and scalable item generation.

- It would be beneficial to have access to more training on prompt engineering, focusing on what works best and what is most and least efficient. It could help SMEs maximise the AI's potential.

Domain-specific knowledge

- Allow users to upload domain-specific knowledge and customise the system by training it with user-provided data (available in Copilot 2.0).
- Enhance the tool's functionality to support the generation of graphs and enable it to read and interpret graphical content.

User Experience

- Improve the platform's user interface and navigation to make it more intuitive, user-friendly and interactive, allowing for smoother interactions between the user and the tool.
- Enable a version control feature to allow SMEs to track changes, compare different iterations of questions and revert to previous versions if necessary.
- It would be beneficial to be able to save the questions, along with the corresponding justifications, as a Word or PDF document. Having documents readily available outside of the system interface would improve content accessibility. It would reduce the risk of errors or miscommunication caused by inconsistent methods of saving and sharing content.

Formatting

- Enhance functionality for formatting to ensure it matches subject-specific standards and SQA's house style.

It is worth noting that Surpass is actively developing Copilot 2.0, which will expand its capabilities to full item generation and potentially resolve some of the concerns raised in this report. The new version will use the latest AI technology (GPT 4o model), which is believed to improve the quality of generated content and will introduce a number of new features such as:

- Possibility of uploading domain knowledge.
- Auto-completion: Copilot will make suggestions to complete partially authored items.
- Customer-controlled training data and copyright control: the option to choose whether to utilise public knowledge data or limit it to organisations' own provided training/ domain data.

Useful resources:

[Surpass Copilot 2.0 functionality video](#)

[Surpass Copilot 2.0 FAQ eBook](#)

5. Proposed next steps

- Trial Copilot v2.0 to evaluate its new functionality and determine if it offers improvements over the previous version.
- Trial ChatGPT Plus for item development. This version allows users to upload domain knowledge and is user-friendly with intuitive navigation.

Appendix 1

Subject Matter Experts (SMEs) evaluation of AI-generated distractors

Paper ID:	
Subject:	
SME:	

Complete the table below with your feedback on AI-generated distractors

1. Reflect on each criterion based on the provided guiding questions
2. Choose the rating that best represents how the criteria have been achieved.

Success Criteria Scale

Scale		Description
1	Poor	Distractors fail to meet expectations, there are significant issues, requiring extensive human intervention
2	Fair	The AI shows some benefits, but there are notable deficiencies in efficiency, quality and accuracy, leading to a less reliable output
3	Good	Distractors meet the basic requirements with a few minor issues. The quality is generally satisfactory with some areas requiring improvement
4	Very good	Distractors are of very good quality, with minor issues that would only require some human intervention. There are clear benefits to the use of AI.
5	Excellent	Distractors are of high quality, with minimal or no issues. The AI exceeds expectations.

Success Criteria	Guiding questions	Feedback	Rating
Quality of distractors	<ul style="list-style-type: none">• Are the distractors plausible and relevant to the questions?• Are there any instances where distractors are misleading or ambiguous?• Are the distractors varied and not repetitive across different questions?• Are there any patterns that reduce the effectiveness of the distractors?• Do the distractors show less/ more bias than those created by SMEs?		

Item difficulty and discrimination	<ul style="list-style-type: none"> • Do the distractors match the intended difficulty level of the questions? • Would they discriminate well between high and low-performing test takers? 		
Time efficiency	<ul style="list-style-type: none"> • Would you need to make significant adjustments to make distractors fit for purpose? • Are there any limitations that could affect scalability? • Would you agree that the use of Surpass Co-pilot could enhance the item development process 		
Overall Quality	<ul style="list-style-type: none"> • How would you rate the quality of AI-generated distractors compared to human-generated ones? • Are you satisfied with the quality and relevance of the AI-generated distractors? • Would you recommend using AI-generated distractors for formative/ summative tests? 		

Appendix 2

SME: Andrew Shield

Higher Physics [2024 Higher Physics Paper 1 Multiple choice Question Paper \(sqa.org.uk\)](https://www.sqa.org.uk)

Prompts and AI-generated distractors

Q1

Prompts:

Ensure all distractors are common arithmetical mistakes made when using the formula $v^2 = u^2 + 2as$.

Ensure answers are given to 2 significant figures.

Ensure answers are not based on selecting the wrong formula.

Ensure answers complete the sentence in the stem that asks the question.

Suggested Distractors:

50 m

This is a common mistake when using the formula $v^2 = u^2 + 2as$. It is likely that the model will generate this as a distractor due to the incorrect use of the formula.

Incorrect use of the formula for distance travelled during acceleration ($v^2 = u^2 + 2as$).

+ Add To Item

10 m

This is another common mistake when using the formula $v^2 = u^2 + 2as$. It is likely that the model will generate this as a distractor due to the incorrect use of the formula.

Incorrect use of the formula for distance travelled during acceleration ($v^2 = u^2 + 2as$).

+ Add To Item

23 m

This answer is based on selecting the wrong formula. The model may generate this as a distractor if it is not well-trained.

Incorrect use of the formula for distance travelled during acceleration.

+ Add To Item

The rationale given for the first two suggested distractors is unclear, and I cannot work out how the AI is coming up with these numbers.

The third suggested distractor has ignored the instruction not to base suggestions on selecting the wrong formula.

None of these are usable distractors.

20 m

This distractor may be chosen by mistakenly using the formula for calculating distance when given initial and final velocities and acceleration. However, this formula is only valid for constant acceleration, and in this scenario the acceleration is not constant. Therefore, this answer is incorrect.

<https://www.physicsclassroom.com/class/1dkin/Lesson-6/Kinematic-Equations>

+ Add To Item

14 m

This distractor may be chosen by using the wrong formula for calculating distance when given initial and final velocities and acceleration. The correct formula to use in this scenario is $d = (v_1 + v_2)/2 * t$, where t is the time taken for the acceleration.

Using this formula with the given values, the distance travelled would be 14 m. However, this formula is not applicable as the time taken for the acceleration is not given in the question, making this answer incorrect.

<https://www.physicsclassroom.com/class/1dkin/Lesson-6/Kinematic-Equations>

+ Add To Item

While the first distractor is possible numerically, it would be suitable but not for the reasoning suggested by the AI. The AI is incorrect in basing its assumptions on the acceleration not being constant – the acceleration is defined as a single value and is therefore constant.

The reasoning behind the second suggestion is invalid, as an alternative way of calculating the response is to first use the data to calculate t and then calculate s . I can generate 14 by making two separate arithmetic mistakes, but it's not one we would choose to use, because it relies on making two mistakes.

27 m

This distractor may be chosen by using the wrong formula for calculating distance when given initial and final velocities and acceleration. The formula $d = \frac{v_f^2 - v_i^2}{2a}$ may be used, which would give the distance travelled as 27 m. However, this formula is not applicable as the acceleration is not constant in this scenario, making this answer incorrect.

<https://www.physicsclassroom.com/class/1dkin/Lesson-6/Kinematic-Equations>

+ Add To Item

This is not a valid rationale for the response suggested. Using the relationship shown gives the correct answer of 17 m. Again, it is making an incorrect statement about acceleration not being constant. I cannot see a way of using the data given to give a value of 27.

Summary for question 1. Most of the AI distractors are unusable, and the one that is usable is for a different reason to the rationale given by the AI. It would be much quicker for the person writing the question to generate distractors and the distractors are more likely to be appropriate ones to use.

Q5

Ensure answers are based on common arithmetic mistakes that candidates may make when using the correct formulae.

Ensure answers are not based on the selection of incorrect formulae.

Ensure answers are given in scientific notation and to two significant figures.

Ensure answers complete the sentence in the stem of the question.

$2.3 \times 10^7 \text{ W}$

This option is incorrect because it is the same value as the given answer but rounded to one significant figure.

The correct answer is given in scientific notation and rounded to two significant figures.

+ Add To Item

$1.5 \times 10^8 \text{ W}$

This option is incorrect because it is the total power delivered from the reservoir to the turbines, not the rate at which water flows.

The correct answer is given in scientific notation and rounded to two significant figures.

+ Add To Item

$8.0 \times 10^7 \text{ W}$

This option is incorrect because it does not take into account the height difference between the reservoir and the turbines.

The correct answer is given in scientific notation and rounded to two significant figures.

+ Add To Item

None of the options have been given in scientific notation as requested.

The first suggested distractor has a rationale that is complete nonsense. It has given an incorrect answer but to two significant figures as requested, and claims that it is the same as the correct response but rounded to 1 significant figure.

The second distractor is also nonsense and I cannot arrive at the quoted number, no matter what I try with the data.

For the third distractor, ignoring the height difference gives a numerical response of $7.5 \times 10^4 \text{ W}$ and not what it is suggesting.

Q7

Answers must be based on applying time dilation formula.

Answers must not be based on choosing an incorrect formula.

Answers must be based on common, likely arithmetical mistakes that candidates will make when calculating the dilated time.

Answers must be given to three significant figures.

Answers must not be randomly generated incorrect numbers, they must be based on using the data given in the question.

20.2 s

This answer is based on a common mistake of forgetting to account for the time dilation effect when calculating the duration of the signal. The correct formula to use in this scenario is $t' = t / (1 - v^2/c^2)$, where t is the time measured on Earth and t' is the time measured on the spacecraft. Plugging in the given values, we get $t' = 20.4$ s, but some candidates may mistakenly use the incorrect formula $t' = t / (1 + v^2/c^2)$ and get a result of 20.2 s.

<https://www.physicsclassroom.com/class/relativity/Lesson-3/Time-Dilation>

+ Add To Item

19.8 s

This answer is based on another common mistake of using the incorrect formula $t' = t / (1 + v^2/c^2)$. Plugging in the given values, we get $t' = 19.8$ s, which is incorrect. This mistake can arise from forgetting to account for the time dilation effect, or from mixing up the equations for time dilation and length contraction.

<https://www.physicsclassroom.com/class/relativity/Lesson-3/Time-Dilation>

+ Add To Item

21.0 s

This answer is based on a mistake of using an incorrect value for the speed of light. Instead of using the value of $c = 3.00 \times 10^8$ m/s, some candidates may mistakenly use a rounded value of 3.01×10^8 m/s, which would result in a time dilation factor of about 1.003 instead of the correct value of 1.002.

Thus, using this incorrect value, the calculated duration of the signal would be 21.0 s instead of the correct answer of 20.4 s.

<https://www.physicsclassroom.com/class/relativity/Lesson-3/Time-Dilation>

Having given the prompt that responses must not be based on choosing an incorrect formula (the required formula is given in the relationships sheet that we give candidates), the AI has generated the first two responses based on ignoring the prompt. It also hasn't stated the correct formula either, in its rationale. Neither distractor would be considered plausible or usable.

The third distractor is based on a completely nonsensical rationale of using 3.01... . This just would not happen (they're given 3.0... in the data sheet). Plus, the speed of light would not round to 3.01, it rounds to 3.00.

None of these distractors is in any way usable. The actual distractors in the question worked and were based on the common arithmetical mistakes that candidates would make when determining the time.

Q9

Ensure answers are in the format of I only, or I and III only, or I, II and III.

Ensure answers cover a range of incorrect and correct responses.

II and III only

Statement II is incorrect because the peak frequency of radiation emitted is greatest for star X, not star W. Statement III is correct as shown by the graph, but statement I is incorrect because it cannot be determined from the graph which star is hotter.

The graph only shows the relationship between energy emitted per second per unit area and wavelength of radiation, but does not provide any information about the temperature or hotness of the stars. Therefore, statement I cannot be determined from the given information.

+ Add To Item

I, II and III

All three statements are incorrect. Statement I cannot be determined from the given information, statement II is incorrect because the peak frequency of radiation emitted is greatest for star X, and statement III is incorrect because star X emits more energy per second per unit area than star Y.

The graph only shows the relationship between energy emitted per second per unit area and wavelength of radiation, but does not provide any information about the temperature or hotness of the stars. Therefore, statement I cannot be determined from the given information. Statement II is incorrect because the peak frequency of radiation emitted is greatest for star X, not star W. And statement III is incorrect because star X emits more energy per second per unit area than star Y.

+ Add To Item

I and II only

Statement I is correct as shown by the graph, but statement II is incorrect because the peak frequency of radiation emitted is greatest for star X, not star W.

The graph shows that star Z emits more energy per second per unit area than star W, indicating that star Z is hotter than star W. Therefore, statement I is correct. However, statement II is incorrect because the peak frequency of radiation emitted is greatest for star X, not star W.

+ Add To Item

II only

Statement II is incorrect because the peak frequency of radiation emitted is greatest for star X, not star W. This can be seen on the graph where the peak frequency for star X is at a higher point than the peak frequency for star W.

<https://www.example.com/graph>

+ Add To Item

II and III only

Statement II is incorrect because the peak frequency of radiation emitted is greatest for star X, not star W. This can be seen on the graph where the peak frequency for star X is at a higher point than the peak frequency for star W. Statement III is incorrect because star X actually emits more energy per second per unit area than star Y. This can also be seen on the graph where the curve for star X is higher than the curve for star Y.

<https://www.example.com/graph>

+ Add To Item

II, III, and IV

Statement II is incorrect because the peak frequency of radiation emitted is greatest for star X, not star W. This can be seen on the graph where the peak frequency for star X is at a higher point than the peak frequency for star W. Statement III is incorrect because star X actually emits more energy per second per unit area than star Y. This can also be seen on the graph where the curve for star X is higher than the curve for star Y. Statement IV is incorrect because star X is not actually the hottest star. According to the graph, star Z has the highest peak frequency and therefore, is the hottest star.

<https://www.example.com/graph>

Whilst I have managed to get the AI to give the correct format for responses, rather than it generating random statements of its own, and the suggestions are mostly usable. The rationales for each are not correct. For the last option, it has introduced an option of statement IV, when there are only three. So that one is completely unusable.

Statements I and III are correct, statement II is incorrect.

Suggestion 1: II and III only is a usable distractor, but the peak frequency is actually greatest for star Z and not star X as claimed by the AI. Statement III is correct and not incorrect as stated by the AI, it is determined from the graph and a bit of knowledge.

Suggestion 2: a usable option but the AI thinks all three statements are incorrect, which isn't correct, as I and III are correct.

Suggestion 3: a usable distractor but not for the incorrect rationale given by the AI. The greater the peak wavelength the lower the temperature.

The other suggestions (apart from the one with statement IV, which is unusable) fall under similar categories – a usable distractor but not for the reasons given by the AI.

Q10

Ensure all distractors are particles covered in the SQA Higher Physics course specification.

Ensure all distractors cannot be classified as a fermion.

Electron

The electron is a fermion, but it is already included in the forbidden list. Choosing it as a distractor would be incorrect and would not fulfill the requirements.

SQA Higher Physics course specification

+ Add To Item

Proton

The proton is a hadron and cannot be classified as a fermion. It is also already included in the forbidden list.

SQA Higher Physics course specification

+ Add To Item

Photon

The photon is a boson and cannot be classified as a fermion. It is also unrelated to the question and would be a good distractor.

SQA Higher Physics course specification

+ Add To Item

Neutron

The neutron is a hadron and cannot be classified as a fermion. It is also unrelated to the question and would be a good distractor.

SQA Higher Physics course specification

Only the 'photon' distractor is usable and for the reason stated.

The suggestion of electron isn't appropriate – the AI has ignored the instruction that it cannot be classified as a fermion (and the rationale actually states it is one).

The suggestions of proton and neutron are also incorrect. Whilst protons and neutrons are hadrons, they are also fermions. The AI has suggested a further three correct responses and only one potential distractor.

Q12

Ensure distractors are either from misunderstanding how beta decay affects the atomic number and mass number or from confusing it with alpha decay and misunderstanding the effect on the atomic number and mass number.

Ensure the element symbol is correct for the atomic number chosen to account for the decay.

Ensure options are expressed in the correct chemical format.



Surpass Copilot was unable to generate any content from your request. Check the contents of your item, provide any additional useful information via the 'Additional Context' box below and try again.

I could not get the AI to generate distractors for this type of question.

Q14

Ensure all distractors are things that could be altered within the experimental set up or the experimental procedure.

Ensure no responses involve changing the light meter.

Ensure responses are specific and not generic.

Using a different light meter

Using a different light meter would not necessarily reduce the systematic uncertainty in the experiment. The light meter used may not be the source of the uncertainty, and changing it could introduce new sources of error.

<https://www.sciencedirect.com/science/article/pii/S0030402612009305>

+ Add To Item

Increasing the distance between the lamp and the light meter

Increasing the distance between the lamp and the light meter may actually increase the systematic uncertainty in the experiment. This could be due to factors such as changes in the intensity of the light or the accuracy of the measurements.

<https://www.nature.com/articles/srep06558>

+ Add To Item

Using a different type of light source

Using a different type of light source may not necessarily reduce the systematic uncertainty in the experiment. The type of light used may not be the source of the uncertainty and changing it could introduce new sources of error.

<https://www.sciencedirect.com/science/article/pii/S0030402612009305>

+ Add To Item

These distractors were generated using just the first prompt. All are unusable and indeed based on spurious arguments for being incorrect.

Using a different type of light filter

Altering the type of light filter used could potentially change the irradiance readings and reduce the systematic uncertainty in the experiment.

<https://www.sciencebuddies.org/science-fair-projects/references/controls-and-variables>

+ Add To Item

Placing the light meter at a different angle

Changing the angle at which the light meter is placed could potentially affect the irradiance readings and reduce the systematic uncertainty in the experiment.

<https://www.sciencebuddies.org/science-fair-projects/references/controls-and-variables>

+ Add To Item

Using a different type of data collection software

Changing the data collection software could potentially affect the accuracy and precision of the irradiance readings and reduce the systematic uncertainty in the experiment.

<https://www.sciencebuddies.org/science-fair-projects/references/controls-and-variables>

+ Add To Item

These three distractors were generated by adding further prompts. Again, none are usable and all are based on spurious arguments. For example, suggesting a different light filter when there isn't one in the experiment. It appears that the AI has been unable to determine what a systematic uncertainty is.

We would use none of the suggestions and end up writing our own.

Q15

Ensure answers are in the format of I only, or I and III only, or I, II and III.

Ensure answers cover a range of incorrect and correct responses.

I and III only

Statement I is correct, but statement III is incorrect. Coherent waves do not necessarily have the same speed.

<https://www.physicsclassroom.com/class/waves/Lesson-4/Coherent-Sources>

+ Add To Item

II and III only

Statement II is correct, but statement III is incorrect. Coherent waves do not necessarily have the same speed.

<https://www.physicsclassroom.com/class/waves/Lesson-4/Coherent-Sources>

+ Add To Item

II only

Statement II is correct, but statements I and III are incorrect. Coherent waves do not necessarily have a constant phase relationship or the same speed.

<https://www.physicsclassroom.com/class/waves/Lesson-4/Coherent-Sources>

Whilst I've managed to get the system to generate the correct style of response, and the options could be used, the AI rationale behind each is incorrect. All three statements in the question are correct and indeed the strict definition of coherence is that the waves have a constant phase relationship (which means they have the same frequency and speed). The AI 'thinks' that some of the facts are not true.

Q17

Ensure all distractors are given in scientific notation.

Ensure all distractors are given to three significant figures.

Ensure all distractors are based on common arithmetical errors when calculating the speed.

Ensure all distractors are based on using a refractive index for diamond of 2.42 and a refractive index for air of 1.00.

Ensure distractors are not based on selecting an incorrect formula.

Ensure distractors are not based on rounding errors.

$$6.16 \times 10^8 \text{ m s}^{-1}$$

This answer is obtained by multiplying the refractive index of diamond (2.42) with the speed of light in a vacuum ($3.00 \times 10^8 \text{ m s}^{-1}$). This is a common error as the speed of light in a medium is always less than its speed in a vacuum.

<https://www.sciencedirect.com/science/article/abs/pii/S0375960197005238>

+ Add To Item

$$1.23 \times 10^8 \text{ m s}^{-1}$$

This answer is obtained by rounding the correct answer to one significant figure. This is a common rounding error that many students make when calculating speeds.

https://www.jstor.org/stable/30007408?seq=1#metadata_info_tab_contents

+ Add To Item

$$1.23 \times 10^{14} \text{ Hz}$$

This answer is obtained by incorrectly converting the frequency in air to the speed in diamond using the refractive index. This is a common mistake as the refractive index only affects the speed of light, not its frequency.

https://www.researchgate.net/publication/221785487_The_speed_of_light_in_a_diamond

+ Add To Item

None of these distractors are usable and indeed the last one isn't even a speed, it's a frequency. The AI has ignored the instruction not to base answers on rounding errors. It has also failed to present the options in scientific notation, as requested.

Q18

Ensure all distractors are given to two significant figures.

Ensure all distractors are based only on using the values given in the circuit diagram.

Ensure all distractors are based on making errors in combining the resistances in the circuit.

Ensure that responses are based on the correct direction of current flow - that current is a flow of electrons from negative to positive.

5 V

This distractor option is based on combining the resistances in the circuit incorrectly. It is half of the correct answer and may be chosen by a student who mistakenly adds the resistances instead of using the formula for voltage in a series circuit.

Ohm's Law states that the voltage in a series circuit is equal to the sum of the individual voltage drops across each component.

+ Add To Item

15 V

This distractor option is based on combining the resistances in the circuit incorrectly. It is one and a half times the correct answer and may be chosen by a student who mistakenly multiplies the resistances instead of using the formula for voltage in a series circuit.

Ohm's Law states that the voltage in a series circuit is equal to the sum of the individual voltage drops across each component.

+ Add To Item

-10 V

This distractor option is based on a common error of forgetting to account for the direction of current flow. In this circuit, the current is flowing from the positive to the negative terminal, so the voltage should be positive. This distractor may be chosen by a student who forgets to consider the direction of current flow.

Voltage is a measure of the potential difference between two points in a circuit and is always positive. When calculating voltage drops, the direction of current flow must be taken into account.

I added the final prompt after generating these three distractors, as the AI had made a fundamental error about electric current. None of these are usable and it would appear that the AI isn't able to pick up what is in the diagram, as it is a mixed series and parallel circuit.

After adding the final prompt, it generated the following:

12 V

This answer is incorrect because it is the sum of all the resistances in the circuit, which would only be the voltage if the resistances were in a series circuit. However, in this circuit, the resistances are in parallel, so the voltage is divided among each resistance.

Ohm's Law and Kirchhoff's Laws

+ Add To Item

8 V

This answer is incorrect because it does not take into account the direction of current flow. In this circuit, the current flow is from negative to positive, so the voltage would be lower at the end with the negative terminal.

Ohm's Law and Kirchhoff's Laws

+ Add To Item

20 V

This answer is incorrect because it is the sum of all the voltages in the circuit, including the voltage drop across the resistors. However, in this circuit, the voltage drop across each resistor would be different due to the resistances being in parallel.

Ohm's Law and Kirchhoff's Laws

There is still nothing usable here, and all suggestions would be rejected for being completely implausible or not being a response that can be achieved using the data in the question.

Appendix 3

Examples of prompts and AI-generated distractors for ESOL questions

National 5 ESOL and Skills for English B1:

- Generate distractors that are at CEFR (Common European Framework of Reference) B2/B1 level
- Distractors should be level-appropriate and test reading comprehension at CEFR B2 Level
- CEFR B2 overall level description: test takers can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.
- Test takers should have an opportunity to demonstrate the following skills (B2):
 - understand detailed language written in English
 - identify the overall purpose, main points and aspects of detail
 - identify how vocabulary, language features and text structures are used to convey meaning
 - identify clearly expressed opinions or attitudes
- Distractors must be clearly incorrect but plausible based on the information provided in the text.
- Distractors should be based on misinterpretations or incorrect inferences of the text that are appropriate at CEFR B2/B1 level
- Distractors should be rooted in the text and should not rely on information outside of the passage
- Distractors can present information that is the opposite of what is stated in the passage, as long as it is sufficiently challenging for B1/B2 level
- All options should be of similar length and style
- In your justifications, explain clearly why distractors are incorrect and a weaker test taker may choose this distractor

Examples of ESOL distractors and justifications (with prompts)

Example 1

Question (refer to paragraph 3-5): The sisters think that responsibility for the problem lies with...

2 Add your answer options and select the correct answers

Answer options

	Correct	Move	Delete
Only the toy retailers are responsible for the gender division of children's toys	<input type="radio"/>	+	x
everyone	<input checked="" type="radio"/>	+	x
The sisters received widespread support and recognition for their first targeted campaign	<input type="radio"/>	+	x
The sisters believe that only boys are affected by gender stereotypes	<input type="radio"/>	+	x

3 Add media to your item (Optional)

Select Media

4 Add text and resources to your item (Optional)

Add To Item

Only the toy retailers are responsible for the gender division of children's toys

This option is not correct because the sisters believe that everyone is responsible for the problem, not just the toy retailers. This option is also incorrect because it does not reflect the sisters' perspective that gender stereotyping is a societal issue and not solely the fault of retailers.

Paragraph 3-5

Added To Item

The sisters received widespread support and recognition for their first targeted campaign

This option is not correct because the sisters mention facing challenges and backlash while starting their campaign, not widespread support. This option is also incorrect because it does not accurately reflect the information provided in the passage.

Paragraph 4

Added To Item

The sisters believe that only boys are affected by gender stereotypes

This option is not correct because the sisters mention that they believe both boys and girls are affected by gender stereotypes, but girls are the main victims. This option is also incorrect because it does not accurately reflect the sisters' perspective.

Paragraph 8

Added To Item

Surpass Copilot misunderstood the task and provided random distractors, not answering the question.

Original question

15. The sisters think that responsibility for the problem lies with

- A manufacturers ☐
- B everyone ☐
- C families. ☐

Example 2

Name: ESOL Nat5 Reading - Question 16 **Status:** Draft **Mark:** 1

women, where some pop star or model or whoever, is what we should all aspire to. But while both women agree that boys, too, can be badly affected by the roles on offer, they see girls as the main victims. "We're going backwards", says Emma, "and it's time to start moving in the other direction."

* Stereotyping: a fixed, over-generalised belief about a particular group or class of people.

Choose the correct answer for each question (refer to paragraphs 3–5).

Question: Some people in the media ...

2 Add your answer options and select the correct answers

Answer options

	Correct	Move	Delete
challenged the sisters	<input checked="" type="radio"/>	+	×
thought the sisters were too radical	<input type="radio"/>	+	×
believed the sisters' campaign was unsuccessful	<input type="radio"/>	+	×
were critical of the sisters' decision to start the Pinkstinks campaign	<input type="radio"/>	+	×

3 Add media to your item (Optional)

Surpass Copilot

thought the sisters were too radical

This option is incorrect because the media did not think the sisters were too radical. In paragraph 5, it is stated that they were featured on television and in newspapers around the world, showing that their campaign was widely accepted and supported.

Pinkstinks, a campaign they set up four years ago to raise awareness of what they say is gender stereotyping that damages children. This week it won a sponsored award from a popular British website for mothers, for promoting body confidence in children. The sisters say they are thrilled, partly because they thought they were too radical for a mainstream award.

Added To Item

believed the sisters' campaign was unsuccessful

This option is incorrect because the media did not believe the sisters' campaign was unsuccessful. In paragraph 5, it is stated that two large department stores responded quickly to criticism, showing that their campaign had an impact.

Two volunteers have been recruited to keep an eye on social media, but otherwise this is it; two working mothers campaigning in the evenings. But they have influence. They have followings on social media and two large department stores have responded quickly to criticism, removing a "girls" label from a set of pink toy figures and a "boys" label from a science kit.

Added To Item

were critical of the sisters' decision to start the Pinkstinks campaign

This option is incorrect because the media were not critical of the sisters' decision to start the Pinkstinks campaign. In paragraph 4, it is stated that they were featured on breakfast television and in newspapers around the world, showing that their campaign was widely accepted and supported.

Their first targeted campaign, in December 2009, attacked the pink/blue colour-coding system used by one popular toyshop. They quickly found themselves on breakfast television and in newspapers around the world.

Added To Item

Distractors are of better quality, but would still require adjustments. They are potentially too open to interpretation and could be deemed too ambiguous at this level

Original question:

16. Some people in the media

A challenged the sisters

☐

B ignored the sisters

☐

C supported the sisters.

☐

Example 3

Name: ESOL Nat5 Reading - Question 18

Status: Draft

Mark

1

Myskow confronted them on TV dressed in pink.

5. The sisters say neither could have done it on her own. "You've got to be so strong to use your voice", says Abi. The sisters run Pinkstinks alongside their day jobs. Emma works for a health research company and Abi is a film-maker, working mainly for charities. Two volunteers have been recruited to keep an eye on social media, but otherwise this is it: two working mothers campaigning in the evenings. But they have influence. They have followings on social media and two large department stores have responded quickly to criticism, removing a "girls" label from a set of pink toy figures and a "boys" label from a science kit.

Finish the sentence: Paragraph 5 tells us that Pinkstinks

+

2 Add your answer options and select the correct answers

Answer options

	Correct	Move	Delete
is powerful despite having few staff.	<input checked="" type="radio"/>		
is understaffed and lacking resources.	<input type="radio"/>		
is unable to make a significant impact due to their small size.	<input type="radio"/>		
is struggling to gain recognition and support.	<input type="radio"/>		
+			

3 Add media to your item (Optional)

Select Media

Surpass Copilot

is understaffed and lacking resources.

This option is incorrect because the article clearly states that the campaign has two volunteers and two working mothers running it in the evenings, which suggests that they have a small but dedicated team. This may be a plausible distractor for a weaker candidate who may not have fully understood the meaning of the word 'influence' in the correct answer and assumed that the campaign is successful because it has many staff members and resources.

Paragraph 5

Added To Item

is unable to make a significant impact due to their small size.

This option is incorrect because the article mentions that the campaign has a following on social media and has successfully influenced two large department stores to change their labeling. This suggests that the campaign has made a significant impact despite having few staff members. This may be a plausible distractor for a weaker candidate who may have not fully understood the meaning of 'influence' and assumed that the campaign has not had any impact.

Paragraph 5

Added To Item

is struggling to gain recognition and support.

This option is incorrect because the article mentions that the campaign has won a sponsored award and has a following on social media. This suggests that they have gained recognition and support for their cause. This may be a plausible distractor for a weaker candidate who may have not fully understood the meaning of 'influence' and assumed that the campaign is struggling to gain recognition and support.

Paragraph 5

Added To Item

Additional Context/Instruction:

Distractors 2 and 4 are relatively good but would require adjustments. Justifications provided are rather superficial and focus on understanding of one word only.

Original question:

18. Paragraph 5 tells us that Pinkstinks

- A benefits from the sisters' day jobs
B sells its material in famous shops
C is powerful despite having few staff.

Example 4 (short text from Skills for English B1 Reading practice test)

Read the text and answer the question that follows:

Not satisfied

I chose this printer because it was cheaper than most of the others online but the amount I saved really wasn't worth all the stress! It took me ages to connect the printer to my wifi. The online instructions are useless - it's like reading a foreign language. And even when it was finally installed, I found that although the quality of the printing is fine, the ink runs out really quickly and costs a lot to replace.

Why is the writer complaining?

2 Add your answer options and select the correct answers

Answer options

printer isn't good enough value for money.

Correct Move



3 Add media to your item (Optional)

Select Media

Suggested Distractors:

printer is not compatible with wifi

This distractor is incorrect because the text clearly states that the printer was eventually connected to a lot of stress. A learner may choose this distractor if they did not pay close attention to the specific details of the text.

The text states, 'It took me ages to connect the printer to my wifi.'

+ Add To Item

the printer is not user-friendly

This distractor is incorrect because although the writer does mention stress and difficulty in connecting the printer, they also mention that the online instructions were useless and the quality of printing is fine. A learner may choose this distractor if they did not pay close attention to all the details in the text.

The text states, 'the online instructions are useless - it's like reading a foreign language' and 'although the quality of the printing is fine.'

+ Add To Item

the printer is too expensive

This distractor is incorrect because the writer mentions that the printer was cheaper than most others but the amount saved was not worth the stress. A learner may choose this distractor if they only focus on the cost in the text.

The text states, 'I chose this printer because it was cheaper than most of the others online but the amount I saved really wasn't worth all the stress!'

+ Add To Item

Some generated distractors could be deemed correct and are open to interpretation. For example 'the printer is not user friendly' The justification for this is a bit nonsensical. The word 'compatible' in the first distractor is above level.

Original question:

The writer is complaining because the...

- a. instructions for the printer aren't in English.
- b. ink for the printer is difficult to find online.
- c. printer isn't good enough value for money.

Example 5 (short text from Skills for English B1 Reading practice test)

Have you got a few spare hours a week?
 If so, why not join our local team of volunteers who provide support to older people living alone?
 A typical visit might include tidying up, preparing a snack or doing a bit of shopping (a car is useful but not necessary).
 Above all, we need kind people who will be available to listen, chat and take an interest in people's lives.
 Please contact us on 07699 228253.

Question: It is essential that volunteers...

+

2 Add your answer options and select the correct answers

Answer options

	Correct	Move	Delete
are friendly	<input checked="" type="radio"/>	+	x
need to work for at least 3 hours a week	<input type="radio"/>	+	x
must have a car to volunteer	<input type="radio"/>	+	x
need to have experience	<input type="radio"/>	+	x

+

3 Add media to your item (Optional)

Select Media

4 Add tags and resources to your item (Optional)

This option is incorrect because the text does not mention any requirement for a background check.

The text mentions: "Please contact us on 07699 228253."

+ Add To Item

need to have experience

This distractor is incorrect because the text clearly states that no previous experience is necessary. A learner who is not at B1 level may choose this distractor because they may not have understood the meaning of the word 'kind' and may think that experience is required to be a volunteer.

CEFR B1: Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.

✓ Added To Item

need to work for at least 3 hours a week

This distractor is incorrect because the text states that only a few spare hours a week are needed, not a minimum of 3 hours. A learner who is not at B1 level may choose this distractor because they may not have understood the concept of 'a few' and may think that a specific number of hours is required to volunteer.

CEFR B1: Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.

✓ Added To Item

must have a car to volunteer

This distractor is incorrect because the text states that a car is useful but not necessary. A learner who is not at B1 level may choose this distractor because they may not have understood the meaning of 'useful' and may think that a car is a requirement to volunteer.

CEFR B1: Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.

✓ Added To Item

Additional Context/Instruction:

Distractor 2 matches the distractor in the original question and distractor 3 could be deemed fit for purpose. Distractor 1 is not really rooted in the text and would not be sufficiently challenging for this level. Justifications for distractors: 'at least 3 hours' and 'have experience' are not very convincing because they fail to reflect the thought process expected from a B1 candidate when approaching this question.

Original question:

It is essential that volunteers...

- a. are friendly. b. can cook well. c. own a car.