
Comparative Study of GenAI (ChatGPT) vs. Human in Generating Multiple Choice Questions Based on the PIRLS Reading Assessment Framework

Speakers: Prof Samuel Chu Kai Wah &
Miss Chloe Lam Yu Yan

Hong Kong Metropolitan University

INTRODUCTION

Reading Battles 2.0: Gamified Learning Platform, 2014-2024



*Samuel Kai Wah Chu
Department of Health Sciences,
School of Nursing and Health Sciences
Hong Kong Metropolitan University*

*Reading Battle is invented by
Prof. Samuel Chu Kai Wah*

Promoting **reading comprehension skill** among students through an engaging **gamified** approach.

Multiple-choice questions (MCQs) aligned with the **PIRLS** reading assessment framework.

Over **10 years** of success in enhancing literacy for Hong Kong student.

Quality Education Fund: Knowledge Overlord - A self-sustaining AI game-based online platform to enhance student's literacy ability and 21st century skills, 1,220,000 USD (Sep 2023 - Aug 2028)

Introduction to KO

- A self -sustaining AI game-based learning platform
- Four pillars of literacy: **reading, writing, questioning, and constructive criticism.**



Quality Education Fund: Knowledge Overlord - A self-sustaining AI game-based online platform to enhance student's literacy ability and 21st century skills, 1,220,000 USD (Sep 2023 - Aug 2028)

General Game Flow

200+ islands of different knowledge domains :

- i.e. Physics Island
- i.e. Fiction Island
- Students must participate in four **literacy battles**
- **Scores accumulated as they battle more**
- Become a knowledge overlord of these islands!



Comparing GenAI and Human-Generated MCQs based on the PIRLS Framework

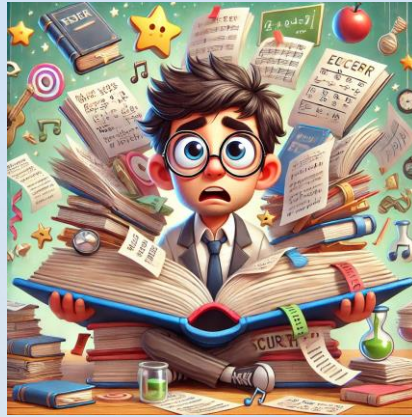
Authors:

- | | | |
|--------------------------|-----------------------------------|-------------------------|
| • Lam, Yu Yan | Hong Kong Metropolitan University | yuylam@hkmu.edu.hk |
| • Chu, Samuel Kai Wah | Hong Kong Metropolitan University | skwchu@hkmu.edu.hk |
| • Ong, Elsie Li Chen | Hong Kong Metropolitan University | eong@hkmu.edu.hk |
| • Suen, Winnie Wing Lam | Hong Kong Metropolitan University | wlsuen@hkmu.edu.hk |
| • Xu, Lingran | Hong Kong Metropolitan University | lixu@hkmu.edu.hk |
| • Lam, Lavender Chin Lui | University of Hong Kong | lavlcl@connect.hku.hk |
| • Wong, Scarlett Man Yu | University of Hong Kong | scar1022@connect.hku.hk |

Institution: The Hong Kong Metropolitan University, Hong Kong S.A.R. (China)

MOTIVATION

Challenges in MCQ Creation & Our Approach
in last decade



Loading...



Creating high-quality MCQs was
resource-intensive (human
expertise, time)

Reading Battle 2.0 traditionally
relied on project staffs, students,
and/or volunteers for MCQ
creation.

Also required extensive review by
educators (much effort and limit
availability of educators/ experts)

MCQ Creation Challenges and Our Approach in Action

Training and Creating

Training in the PIRLS framework workshop was offered to:

- *Project staff*
- *Students*
- *Volunteers (primarily older adults)*

before creation of questions

Approx: 3 weeks

Careful Reviewing and Selection

- *Questions were submitted to the competition/ review panels*
- *Good- formulated questions were reviewed/ refined by a research team.*

Approx: 2 weeks

Launching to the database

- *The finalized questions were added to the e-quiz database in the cloud*

Approx: 1.5 weeks



From training to implementation, the process could take weeks to months!

Introduction, Background, Motivation of this study

Literature Review

- The PIRLS Framework
- Generative AI for Question Generation
- Comparative Performance in 3 studies

Research Question

Schematic diagram of the study design

Prompt design of the study

A Simple blinded test:

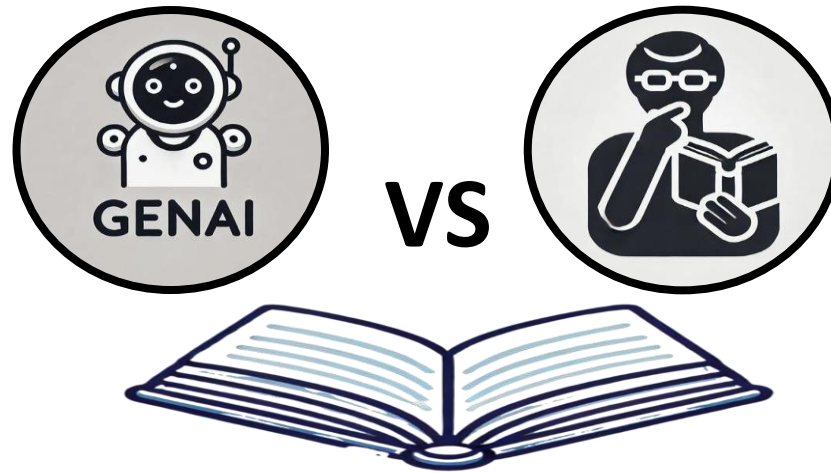
Guess who constructed it?

Assessment Result

Table 1: Identification Accuracy

Table 2: Wilcoxon rank sum tests

Comparing GenAI and Human-Generated MCQs based on the PIRLS Framework



**Introduction, Background, Motivation
Of this study**

Literature Review

- **The PIRLS Framework**
- Generative AI for Question Generation
- Comparative Performance in 3 studies

Research Question

Schematic diagram of the study design

Prompt design of the study

A Simple blinded test:

Guess who constructed it?

Assessment Result

Table 1: Identification Accuracy

Table 2: Wilcoxon rank sum tests

The PIRLS Framework

- **Full name:**
The **P**rogress in **I**nternational
Readng **L**iteracy **S**tudy
- A valuable model for measuring reading comprehension skills in children **globally**.
- A **blueprint recipe** in this study:
 - Guide both human and GenAI in creating high-quality, PIRLS-aligned Multiple Choice Questions



Level 4: Evaluate & Examine Critically

Critical evaluation and forming judgments based on content.



Level 3: Interpret & Integrate ideas

Interpreting ideas and also integrating informations from different parts of the text



Level 2: Make Inferences

Making some simple reasoning/ inferences from the text



Level 1: Retrieve Information

Retrieve basic facts and details.

**Introduction, Background, Motivation
Of this study**

Literature Review

- The PIRLS Framework
- **Comparative Performance in 3 studies**

Research Question

Schematic diagram of the study design

Prompt design of the study

A Simple blinded test:

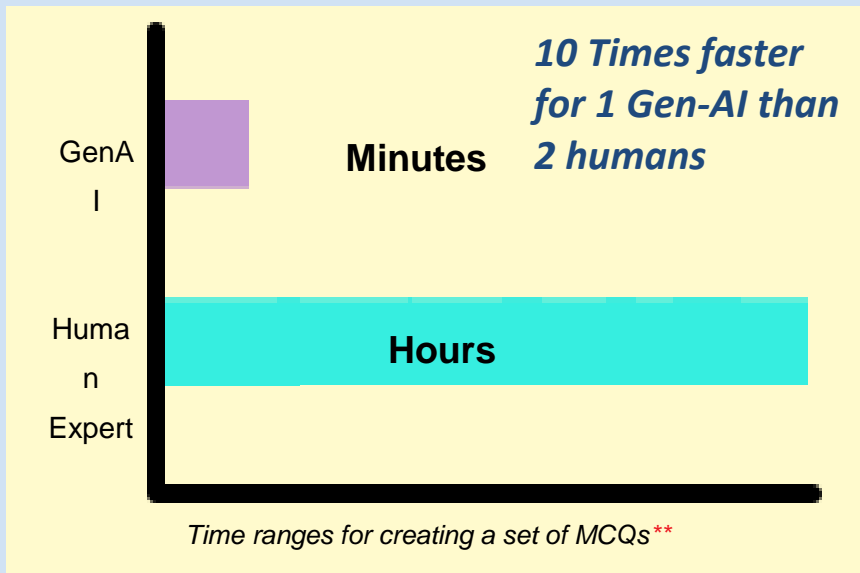
Guess who constructed it?

Assessment Result

Table 1: Identification Accuracy

Table 2: Wilcoxon rank sum tests

Comparative Performance in Question Generation -Time



***The human time range is an approximation or an average across experts*

• Time Performance of Generative AI (Cheung et al., 2023):

- Time required for creating 50 questions
 - ChatGPT: 20 minutes 25 seconds
 - Two human examiners: 211 minutes 33 seconds
- Even experienced examiners required more than **ten minutes** to prepare one question on average

Comparative Performance in Question Generation - Question quality

	Cheung et al. (2023)	Doughty et al. (2024)
Assessment Quality	<ul style="list-style-type: none">• No significant difference in the question quality (i.e. Appropriateness and Clarity)• Only in the relevance category that the A.I. was inferior to humans	<ul style="list-style-type: none">• Comparable clarity and alignment. However, AI questions had multiple correct answers (4.9%).
Other findings	<ul style="list-style-type: none">• AI reduces workload of examiners significantly.	<ul style="list-style-type: none">• Strong learning alignment but human oversight needed for quality.

The current study attempted to explore the power of this GenAI tool further with alignment to the PIRLS assessment framework and in a younger age group.

**Introduction, Background, Motivation
Of this study**

Literature Review

- The PIRLS Framework
- Comparative Performance in 3 studies

Research Question

Schematic diagram of the study design

Prompt design of the study

A Simple blinded test:

Guess who constructed it?

Assessment Result

Table 1: Identification Accuracy

Table 2: Wilcoxon rank sum tests

Research questions

- How do AI-generated MCQs compare to human-crafted ones in terms of quality and alignment with the PIRLS framework?

- Can AI-generated MCQs effectively replace human-created questions for educational assessments?

**Introduction, Background, Motivation
Of this study**

Literature Review

- The PIRLS Framework
- Comparative Performance in 3 studies

Research Question

Schematic diagram of the study design

Prompt design of the study

A Simple blinded test:

Guess who constructed it?

Assessment Result

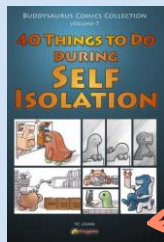
Table 1: Identification Accuracy

Table 2: Wilcoxon rank sum tests

Comparing GenAI and Human-Generated MCQs based on the PIRLS Framework

Schematic diagram of the study design

Pro-Phase: Books selection



Book 1:
Buddysaurus
Comics: 40
Things to Do
during Self
Isolation



Book 2: The Missing Mum



Part 1: Questions Creation

16 human-crafted
MCQs

- MCQs construction (Approx 160 minutes for 2 books)
- MCQs selection (Approx 1 hour)




16 AI-generated
MCQs

- prompt design (3 hours)
- MCQs generation (1 minute for each book)

Part 2: Question Assessment

*Randomized 32 MCQs
(Human and GenAI).*

Panel of Assessors:

- Four experienced reviewers need to determine:  **4**
 - a. Alignment
 - b. Clarity
 - c. Appropriateness
 - d. Suitability
 - e. AI/Human

**Introduction, Background, Motivation
Of this study**

Literature Review

- The PIRLS Framework
- Comparative Performance in 3 studies

Research Question

Schematic diagram of the study design

Prompt design of the study

A Simple blinded test:

Guess who constructed it?

Assessment Result

Table 1: Identification Accuracy

Table 2: Wilcoxon rank sum tests

Prompt Engineering for PIRLS-Aligned MCQs-1

Our Prompt into GPT-4

Based on the following excerpt from a children's book,

[The Missing Mum

The Missing Mum, Irene Yim, Illustrated by Simran Dadlani, Supported by Sam Chu and Helen Chan

Mum suddenly put down her cutlery, turned off the TV in front of the dining table, and solemnly announced to all of us,

"I have a request, I hope you all agree. I am tired of being mum, and I hope to disappear for one day tomorrow. Can you all help me share my chores?"

Dad glimpsed at Ken and May, and asked Mum, "Can you tell us where are you going to disappear to?" "Sure, of course! I am not going anywhere. I will be disappearing into my room."

Ken frowned, and asked her, "Then what are you going to do in there?" "I would be doing nothing, just taking naps, reading, and listening to music. Tomorrow I will be disappearing. Will you guys be alright without me?" "We will try!"

The house was unusually quiet the Sunday morning that Mum disappeared. She was not making breakfast in the kitchen. The washing machine was not working, and nobody was urging Ken and May to wake up.

"Hey, May, wake up and make breakfast. What do you want to eat?" Ken nudged his sister to wake up. "Two fried eggs, help me make it, brother. I want to disappear too!" May whined.

"Nobody is making anything, it is too troublesome. I will go buy some." Dad cried out from the bathroom.

Mum shouted out from her room, "Don't order food from outside. It's unhygienic and not nutritious. Let me cook."

Mum made breakfast quickly and said to everyone: "I'm missing from now on. After you have breakfast, please wash the dishes and then throw the dirty clothes into the washing machine." Mum slammed the door and disappeared into her room.

Thirty minutes later, there was a loud BANG in the kitchen. Mum was startled. She stopped listening to the song by "Andy Lau" by turning off the radio with a snap. She then jumped out of the bed, and put her head out of the door. "What happened?"

"Sister broke the glass and cut her finger!" "WHAT! Come and show me!" Mum stopped. May's cut from bleeding and put a bandage on it. Mum said uneasy, "Ah, let me clean it up!"

Mum quickly cleaned up the kitchen and wiped her hands while saying to everyone: "I am really going to go missing this time, please don't bother me again."

Dad threw the dirty clothes into the washing machine. "Ken, why doesn't this automatic washing machine work by itself?" "I don't know, look at the manual!" "Where is the manual?" "In mum's head!" "Please take it out!"

Ken knocked on mum's door. "Mum, I'm sorry to disturb you, but the washing machine won't work!" She stopped listening to the song from "George Lam" by turning off the radio with a snap. "Ah, let me teach you guys!"

Mum put the washing powder in the machine and turned on the tap. Her hands gently patted the door of the washing machine like she was playing a magic trick. Everyone was surprised and admired their mum's extraordinary strength!

"Mum, you are amazing! Don't go missing, we can't live without you! You can teach us to do housework!" everyone said, hugging their mum. "Yes, mum cannot go missing! There is no such thing as Missing Mum on earth!"

], create 8 multiple-choice questions suitable for 5th graders. Ensure each pair of questions aligns with a different level of the PIRLS reading assessment framework to test a range of comprehension skills. While creating these questions, incorporate the comprehension skills as detailed in the "**Guidance on PIRLS Levels**" and "**A Sample passage and the related examples for 4 PIRLS Levels**". Drawing inspiration from the examples provided, generate 8 MCQs suitable for 5th graders from the book excerpt. Each question should be crafted to align with a distinct level of the PIRLS reading assessment framework, as demonstrated by the examples. Provide four options (A, B, C, D) for each question, clearly identifying the correct answer and offering a brief explanation that connects the correct choice to the excerpt's key ideas or themes.

Guidance on PIRLS Levels:

Level 1: Focus on generating questions that require the retrieval of explicitly stated information from the text.

This level involves identifying and recalling information that is clearly stated in the text, such as the sequence of events, character traits, and making comparisons. The evaluation would focus on whether the question requires the examinee to retrieve specific details directly from the text, aligning with the first element of relevance to the article's content.

Level 2: Create questions that encourage making straightforward inferences from the text, beyond mere retrieval of information.

Questions at this level require the examinee to make direct inferences, such as understanding cause-and-effect relationships or predicting outcomes that are not explicitly stated but can be logically deduced from the text. This involves a deeper engagement with the text and begins to touch upon the second element of the evaluation criteria, where the question tests the examinee's ability to infer and reason beyond the surface details.

Level 3: Develop questions that prompt the interpretation of ideas and integration of information from different parts of the text.

At this level, questions ask the examinee to interpret characters' feelings and behaviours and integrate ideas and information across the text. This requires a comprehensive

Grey color:
Excerpt from
the children's
book, The
Missing Mum

Orange color:
Detailed
instructions
provided to
ChatGPT for
generating 8
MCQs.

Blue color:
Guidance on how
to align the
generated
questions with
PIRLS levels.

Prompt Engineering for PIRLS-Aligned MCQs-2

Prompt for ChatGPT-4.0 Few-Shot Learning Task with Real Examples

*"Based on the following excerpt from a children's book: [insert book excerpt here], create 8 multiple-choice questions suitable for 5th graders. Ensure each pair of questions aligns with a different level of the PIRLS reading assessment framework to test a range of comprehension skills. While creating these questions, incorporate the comprehension skills as detailed in the ****Guidance on PIRLS Levels**** and ****A Sample passage and the real examples for 4 PIRLS Levels****. Drawing inspiration from the examples provided, generate 8 MCQs suitable for 9-year-olds from the book excerpt. Each question should be created to align with a distinct level of the PIRLS reading assessment framework, with two questions on each level. Provide four options (A, B, C, D) for each question, clearly identifying the correct answer and offering a brief explanation that connects the correct choice to the excerpt's key ideas or themes.*

Orange color:
Detailed
instructions
provided to
ChatGPT for
generating 8
MCQs.

Prompt Engineering for PIRLS-Aligned MCQs-3

Purple color:
Few-Shot Learning
(FSL) was applied to
GenAI models, where
ChatGPT was given
explicit MCQ
examples for each
PIRLS level alongside
a sample text excerpt
from the book-
Cracking Coconut

understanding and the ability to synthesise information from different parts of the text, directly engaging with the second element of testing or training reading comprehension abilities.

Level 4: Construct questions that necessitate examining and evaluating the text's content, language, and textual elements in a critical manner.

The highest level of questions challenges examinees to critically evaluate the text's content, language, and textual elements, such as making judgments about worth, desirability, and acceptability or considering how they would react in the place of a character. This level fully embodies the evaluation criteria's second element by requiring an advanced level of comprehension, analysis, and critical thinking.

A Sample Text Excerpt and the real examples for 4 PIRLS Levels

A Sample Text Excerpt:

[Cracking Coconut, Author: Winnie Suen, Illustrator: Winnie Suen, Editor In Chief: Sam Chu and Helen Chan

Cassie went on a trip to the Philippines with her Dad, Carl.

She saw a coconut palm next to her host's farm.

The coconut palm is as tall as her house.

"What can we use coconut for?" Cassie asked her Dad.

Her Dad replied "Coconut is a miracle fruit."

"It can be used for its water, milk, oil and meat."

Cassie's eyes widened and she asked "How does it taste?"

Her Dad replied "It is slightly sweet. Do you want to try one?"

Cassie nodded. Her Dad then brought her to a supermarket.

They purchased a fresh coconut.

Cassie could not wait and grabbed it over to drink.

"Wow, it tastes so nice!" Exclaimed Cassie.

"How about buying more to share with our host?" Cassie asked her Dad.

Carl smiled at Cassie. "Of course. Let's share the joy!"

Example for Level 1:

{

Who Cassie wanted to share more coconut with? (Aligned with PIRLS Level 1: Focus on and Retrieve Explicitly Stated Information)

Her father

Her mother

Her host (Correct Answer) - Explanation: ["How about buying more to share with our host?"]

Cassie asked her Dad."

Her friends

}

Example for Level 2:

{

Where can we possibly find Coconut Palm? (Aligned with PIRLS Level 2: Make Straightforward Inferences)

Philippines (Correct Answer) - Explanation: [Cassie and her Dad found Coconut palm in Philippines.]

Denmark

China

Sweden

}

Example for Level 3:

{ Why did Cassie have her eyes widened? (Aligned with PIRLS Level 3: Interpretation of ideas and integrating of information)

She was amazed by the fact that coconut can be of multiple use. (Correct Answer)

-Explanation: [She learned more about coconut.]

She was amazed by the fact that coconut is sweet.

She was amazed by the fact that she may have a taste of coconut.

She was terrified by the miracle fruit.

}

Example for Level 4:

{ How would you comment Cassie's act of buying more coconuts to share with her host? (Aligned with PIRLS Level 4: Examine and evaluate content, language, and textual elements)

She is caring and thoughtful. (Correct Answer) -Explanation: [She is caring and willing to share happiness with others.]

She is selfish.

She is not caring.

She is not thoughtful.

}

Reply from GPT-4

Based on the provided excerpt from "The Missing Mum," here are eight multiple-choice questions crafted for 5th graders, each aligning with the distinct levels of the PIRLS reading assessment framework:

Level 1: Focus on and Retrieve Explicitly Stated Information

Question 1: What did Mum plan to do for one day?

- A) Go on a vacation

- B) Disappear into her room (Correct Answer)

- C) Visit a friend

- D) Go shopping

- **Explanation:** The text explicitly states that Mum planned to "disappear" into her room for one day.

Question 2: What caused Mum to leave her room the first time?

Purple color:
Real Examples
created by
human experts
for each PIRLS
level.

Green Color:
ChatGPT-4 Output
showing how the AI-
generated multiple-
choice questions
align with the
different PIRLS
levels.

**Introduction, Background, Motivation
Of this study**

Literature Review

- The PIRLS Framework
- Comparative Performance in 3 studies

Research Question

Schematic diagram of the study design

Prompt design of the study

A Simple blinded test:

Guess who constructed it?

Assessment Result

Table 1: Identification Accuracy

Table 2: Wilcoxon rank sum tests

Question Number	Question	Options	Correct Answer	PIRLS Level	Alignment with PIRLS Level (1-10)	Clarity and Specificity (1-10)	Appropriateness (1-10)	Suitability for Specific Age Group (1-10)	Source (Human/AI)
1	What does the author use to represent the concept of Buddysaurus?	A) Buddy + Dinosaur + Us B) Friendship + Time C) Comics + Dinosaurs D) Laughter + Fun	A	1					?
2	What is the full form of "FAIL" ?	A) First attempt in learning B) First attempt in laughing C) First attempt in losing D) Final attempt in learning	A	1					?
3	What is the main mission of the Buddysaurus Comics according to YC Chan?	A) To become famous B) To remind you of your Buddies and inspire you to be better C) To sell comic books	B	1					AI
4	What can we do during self-isolation?	A) Gather and eat with family and friends B) Go to a good restaurant to relax C) Play at a friend's house	D	1					Human
5	How might the Buddysaurus help its readers according to the text?	A) By teaching them about dinosaurs B) By encouraging them to read more C) By inspiring positive changes and better friendships	C	2					AI
6	What do we have to remember to do after eating good food during self-isolation?	A) Watch new movies B) Sleep in bed C) Get enough exercise	C	2					Human
7	What does "sleeping too well" mean?	A) Sleeping just right B) Sleeping too much C) All of the choices	B	2					Human
8	Why does the character feel like belonging to an extinct species before meeting a friend?	A) They actually are a dinosaur. B) They were alone and felt unique. C) They didn't like comics.	B	2					AI
9	Throughout the whole book, how can we be mentally healthy during self-isolation?	A) Connecting with friends and family online B) Eat a lot of fried food C) Sleeping 15 hours a day	A	3					Human
10	What does the interaction between Spiky and Long Neck teach us about friendship?	A) Friends should always give advice. B) Listening is sometimes more important than speaking.	B	3					AI
11	What is NOT a good example of multi-tasking?	A) Organizing your room and listening to music B) Exercising and watching movies C) All of the choices	D	3					Human
12	What does the phrase 'it takes a long time to grow an old friend' suggest about friendships?	A) Friendships form quickly. B) Friendships don't need time to develop. C) The best friendships require patience and time to	C	3					AI
13	How does the Buddysaurus mission aim to address loneliness?	A) By encouraging self-reliance B) By promoting the importance of having friends to tackle challenges together	B	4					AI
14	How do you think a person will feel if forced to self-isolate for a long time?	A) Anxious B) All of the choices C) Unhappy	B	4					Human
15	What is the significance of comparing buddies to Jade and Amber?	A) To show that friendships are rare B) To illustrate that friendships, like precious stones, need to be protected C) To suggest that friendships are expensive	B	4					AI
16	How can we care about someone during self-isolation?	A) Show your love to your family and friends B) All of the choices C) Keep in contact with your family and friends	B	4					Human

Assessment tool - 4 Key Domains for Question Quality Comparison

(adapted from *Cheung et al.,2023*)

- 1. Alignment with PIRLS Levels:** Determine if the question is set at the level stated (PIRLS 1, 2, 3 and 4)
- 2. Clarity and Specificity:** Determine if the question is clear and answerable
- 3. Appropriateness:** Determine if the question is with appropriate length and well-formulated.
- 4. Suitability for Specific Age Group:** Determine whether the question is suitable for assessing 9-year-old students.
 - Numeric scale from 0- 10 ("0" - extremely disagree, "10" - extremely agree)
5. Blinded test of question origin: AI/human constructed?

The Mini Game Time

See whether you could tell the difference between AI-generated and human crafted question!!!!



Scan the QR code to enter the game-Kahoot!

Can you guess the origin of the MCQs?

Q1.

Who broke the glass and was injured?

- A. Mother
- B. May (Correct Ans)
- C. Father
- D. Ken



or



PIRLS Level 1: Information retrieval

Q2.

What did Mum plan to do for one day?

- A. Go on a vacation
- B. Disappear into her room (Correct Ans)
- C. Visit a friend
- D. Go shopping



or



PIRLS Level 1: Information retrieval

Q3.

Why did May and Ken wake up late on Sunday?

- A. Their father did not wake them up.
- B. They wanted to sleep as long as they like.
- C. They forgot to set the alarm.
- D. The person who usually wakes them up was absent that day.(correct)



or



PIRLS Level 2: Simple Inference

Q4.

Why did Mum want to disappear for a day?

- A. She wanted to clean the house in peace.
- B. She wanted to prepare a surprise for the family.
- C. She was feeling tired and needed a break.(correct)
- D. She needed to work on a secret project.



or



PIRLS Level 2: Simple Inference

**Alignment
with PIRLS
Level (1-10)**

**Clarity and
Specificity
(1-10)**

**Appropriateness
(1-10)**

**Suitability for
Specific Age
Group (1-10)**

**Source
(Human/
AI)**

Can you guess the origin of the MCQs?

Q1.

Who broke the glass and was injured?

- A. Mother
- B. May (Correct Ans)
- C. Father
- D. Ken



PIRLS Level 1: Information retrieval

Q2.

What did Mum plan to do for one day?

- A. Go on a vacation
- B. Disappear into her room (Correct Ans)
- C. Visit a friend
- D. Go shopping



PIRLS Level 1: Information retrieval

Q3.

Why did May and Ken wake up late on Sunday?

- A. Their father did not wake them up.
- B. They wanted to sleep as long as they like.
- C. They forgot to set the alarm.
- D. The person who usually wakes them up was absent that day.



or



PIRLS Level 2: Simple Inference

Q4.

Why did Mum want to disappear for a day?

- A. She wanted to clean the house in peace.
- B. She wanted to prepare a surprise for the family.
- C. She was feeling tired and needed a break.
- D. She needed to work on a secret project.



or



PIRLS Level 2: Simple Inference

Alignment
with PIRLS
Level (1-10)

Clarity and
Specificity
(1-10)

Appropriateness
(1-10)

Suitability for
Specific Age
Group (1-10)

Source
(Human/
AI)

Can you guess the origin of the MCQs?

Q1.

Who broke the glass and was injured?

- A. Mother
- B. May (Correct Ans)
- C. Father
- D. Ken



PIRLS Level 1: Information retrieval

Q2.

What did Mum plan to do for one day?

- A. Go on a vacation
- B. Disappear into her room (Correct Ans)
- C. Visit a friend
- D. Go shopping



PIRLS Level 1: Information retrieval

Q3.

Why did May and Ken wake up late on Sunday?

- A. Their father did not wake them up.
- B. They wanted to sleep as long as they like.
- C. They forgot to set the alarm.
- D. The person who usually wakes them up was absent that day.



or



PIRLS Level 2: Simple Inference

Q4.

Why did Mum want to disappear for a day?

- A. She wanted to clean the house in peace.
- B. She wanted to prepare a surprise for the family.
- C. She was feeling tired and needed a break.
- D. She needed to work on a secret project.



or



PIRLS Level 2: Simple Inference

**Alignment
with PIRLS
Level (1-10)**

**Clarity and
Specificity
(1-10)**

**Appropriateness
(1-10)**

**Suitability for
Specific Age
Group (1-10)**

**Source
(Human/
AI)**

Can you guess the origin of the MCQs?

Q1.

Who broke the glass and was injured?

- A. Mother
- B. May (Correct Ans)
- C. Father
- D. Ken



PIRLS Level 1: Information retrieval

Q2.

What did Mum plan to do for one day?

- A. Go on a vacation
- B. Disappear into her room (Correct Ans)
- C. Visit a friend
- D. Go shopping



PIRLS Level 1: Information retrieval

Q3.

Why did May and Ken wake up late on Sunday?

- A. Their father did not wake them up.
- B. They wanted to sleep as long as they like.
- C. They forgot to set the alarm.
- D. The person who usually wakes them up was absent that day.



PIRLS Level 2: Simple Inference

Q4.

Why did Mum want to disappear for a day?

- A. She wanted to clean the house in peace.
- B. She wanted to prepare a surprise for the family.
- C. She was feeling tired and needed a break.
- D. She needed to work on a secret project.



PIRLS Level 2: Simple Inference

**Alignment
with PIRLS
Level (1-10)**

**Clarity and
Specificity
(1-10)**

**Appropriateness
(1-10)**

**Suitability for
Specific Age
Group (1-10)**

**Source
(Human/
AI)**

**Introduction, Background, Motivation
Of this study**

Literature Review

- The PIRLS Framework
- Comparative Performance in 3 studies

Research Question

Schematic diagram of the study design

Prompt design of the study

A Simple blinded test:

Guess who constructed it?

Assessment Result

Table 1: Identification Accuracy

Table 2: Wilcoxon rank sum tests

Comparative performance between human and AI in question creation -1

Assessment Result

Questions created by	Each human assessor evaluated 32 questions by AI/Human Number of Q&As correctly Identified by the 4 assessors	Incorrectly identified by human assessors	Total Assessed Questions	Percentage Correctly Identified
AI (16 Questions)	N(Guess AI AI) 21 questions	N(Guess Human AI) 43 questions	64	32.81%
Human (16 Questions)	N(Guess Human Human) 36 questions	N(Guess AI Human) 28 questions	64	56.25%

Result Table 1:
Identification Accuracy of AI-Generated vs. Human-Generated MCQs

Comparative performance between human and AI in question creation -2

Questions created by	Each human assessor evaluated 32 questions by AI/Human Number of Q&As correctly Identified by the 4 assessors	Incorrectly identified by human assessors	Total Assessed Questions	Percentage Correctly Identified
AI (16 Questions)	N(Guess AI AI) 21 questions	N(Guess Human AI) 43 questions	64	32.81%
Human (16 Questions)	N(Guess Human Human) 36 questions	N(Guess AI Human) 28 questions	64	56.25%

$$\text{Accuracy} = \frac{\text{Correct Identifications (TP + TN)}}{\text{Total Questions Assessed}} = \frac{\text{Correct Identifications (21+ 36)}}{\text{Total Questions Assessed(128)}} = 44.5\%$$

Result Table 1:
Identification Accuracy of AI-Generated vs. Human-Generated MCQs

Comparative performance between human and AI in question creation -2

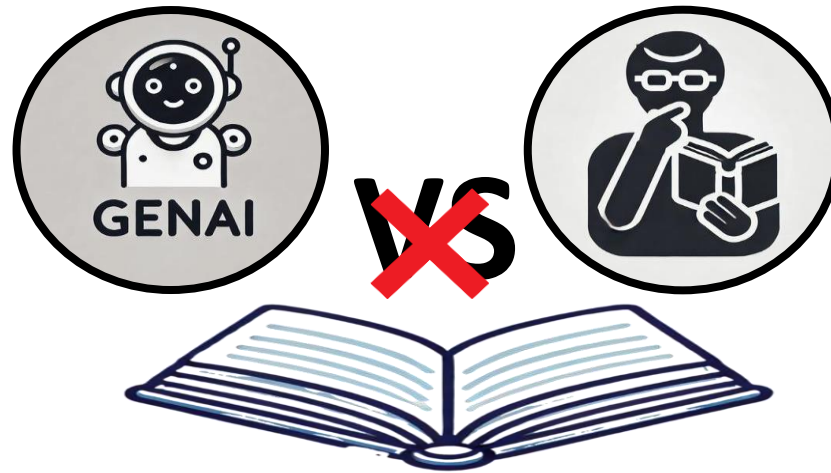
The Wilcoxon rank sum tests comparing AI-generated and Human-generated MCQs across four domains

Assessment Domain	AI Mean Score	Human Mean Score	P
Alignment with PIRLS Level (1-10)	7.98	6.61	0.03485
Clarity and Specificity (1-10)	7.86	6.86	0.04238
Appropriateness (1-10)	7.92	6.56	0.06941
Suitability for Specific Age Group (1-10)	7.81	6.56	0.07876

Research questions

- How do AI-generated MCQs compare to human-crafted ones in terms of quality and alignment with the PIRLS framework?
- Can AI-generated MCQs effectively replace human-created questions for educational assessments?

Comparing GenAI and Human-Generated MCQs based on the PIRLS Framework



Comparing GenAI and Human-Generated MCQs based on the PIRLS Framework

