Evaluation of Large Language Models for Online Assessment Questions

Pedro Malavota Ribeiro

August 2025

1 INTRODUCTION

Objective of the Report

Summarize the research on the use of Large Language Models (LLMs) to enhance the quality of online assessment questions.

Task Context

The team currently works with questions of variable quality. The goal is to test whether LLMs can be reliably used to identify problems, classifycognitive complexity, and suggest improvements.

Report Structure:

This document covers the methodology, the findings from the LLM analysis, and the recommendations based on the results.

2 METHODOLOGY

2.1 Tools Used

- LLMs: GPT-5 (August 7, 2025 timestamp), GPT-5-mini, GPT-5-nano
- Coding Environment: Jupyter Notebook (Python)
- Libraries: pandas, os, json, dotenv, openai.

2.2 Initial Human Analysis

A manual review of 30 questions was conducted according to [1]:

- 5 questions were marked as uncertain (Questions 4, 6, 7, 8 and 23).
- 3 questions were labeled as poorly designed :
    - Ambiguous: Question 6
    - Trivial: Questions 1 and 11
- 3 questions were designed as trick questions to bias test-takers (Questions 2,15, 17).

2.3 Prompt Development

As the most recent LLM models are usually “smarter”, the just-released gpt-5 was chosen for the devolopment of the assessment questions, as it would represent the state-of-the-art in this field.

Preliminary testing was done with gpt-4o-mini, using questions 1–5, using temperature=0. Later, the newly-released gpt-5 was adopted with a time-stamp for more consistent output.

Finally, gpt-5-mini was adopted in batch processing of mandatory functions, then in the open research question, both gpt-5-mini and gpt-5-nano were adopted, due to the lower billing in comparison to gpt-5, and simpler nature of the tasks.

Basic prompting techniques were applied, particularly role assignment [2]. Other practices included tagging for information breakdown and chain-of-thought reasoning, with both positive and negative examples.

For improving models, functions also retrieved a *reasoning* string, to better understand LLM response and address hallucinations.

2.4 Batch Processing of mandatory LLM functions

The *run_models* function was designed to run any of the mandatory LLM function *classify_bloom_level*, *validate_question_quality*, *suggest_question_improvement*, *recommend_format_conversion* and *generate_format_conversion*. It also created a file labeled with the question ID and name of the called function, where it loaded the LLM outputs.

To process all of the 30 samples, gpt-5 was used to run *classify_bloom_level* and *validate_question_quality* with reduced *effort* parameter and models did not return their reasoning, in order to generate less (expensive) tokens. Finally, *suggest_question_improvement*, *recommend_format_conversion* and *generate_format_conversion*

2.5 Open research question – Automated TEXT Scoring

A lightweight methodology for evaluating short-text answers (2–3 sentences) was developed. It consisted in decomposing an answer into logical arguments and assigning a score to each argument based on logical validity and contextual relevance. The evaluation process relied on a Large Language Model (LLM) to assist in argument segmentation and qualitative judgment.

Step 1: Argument Extraction

Given an answer A, extract a set of N arguments:

$$A \rightarrow \{arg_1, arg_2, \dots, arg_N\}.$$

This decomposition is performed by the LLM, which identifies the main claims or propositions within the text.

Step 2: Logical Validity

For each argument arg_i, the LLM assigns a logical validity score $L_i \in \{0, 0.5, 1\}$:

L_i = 1 if argument is logically valid

2/3 if argument is partially valid or inconclusive

0 if argument is invalid or contradictory

Step 3: Contextual Relevance

Each argument is also evaluated for its relevance to the question Q. The LLM assigns a contextual score $C_i \in \{0, 0.5, 1\}$:

C_i = 1 if directly connected to the question

1/2 if indirectly connected

0 if disconnected

Step 4: Argument Scoring

The score of each argument is computed as the product of its logical and contextual values:

$$S_i = L_i \,.\, C_i$$

Step 5: Final Answer Score

The final score for the answer is the normalized mean of argument scores, scaled to the range [0, 10]:

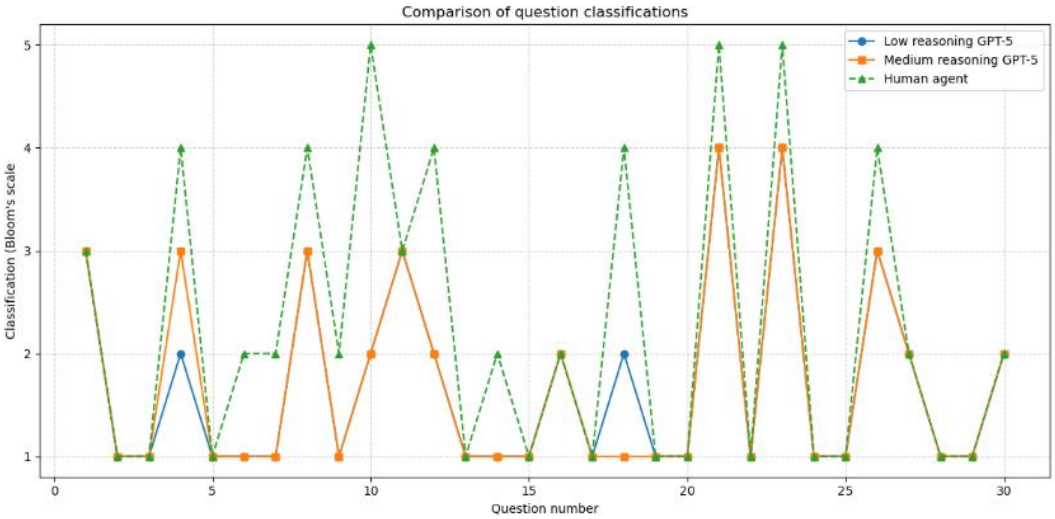
$$Score(A, Q) = 10 \times (1/N)\, \Sigma\, S_i$$

3 RESULTS

3.1 Bloom Level Classifications

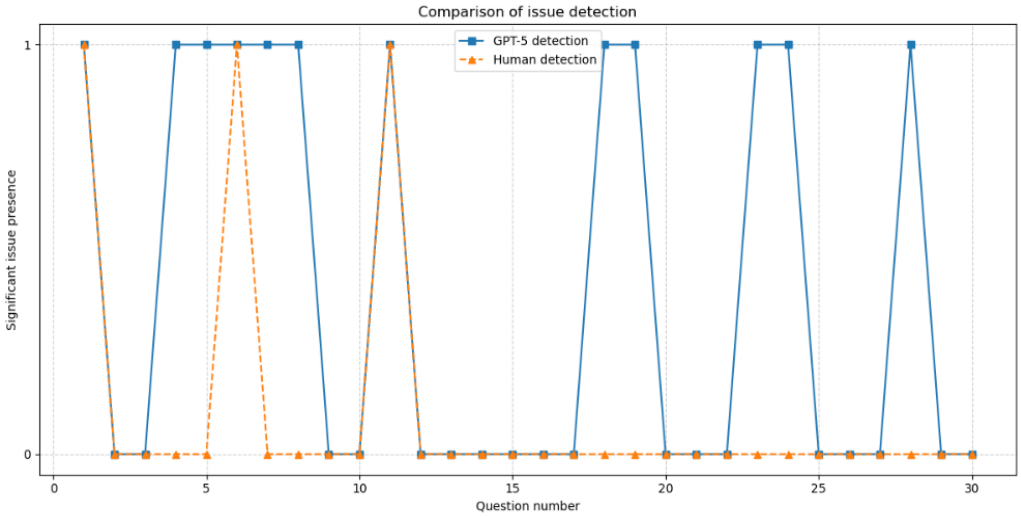
According to Plot 1, human analysis and the LLM model agreed completely on 18 questions: 13 agreements for level 1 classifications, 3 agreements for level 2 and 3 agreements for level 3. 6 questions were labeled with a difference of 1 level, between LLM and human.

The model produced different decisions only at questions 4 and 18. The highest level of thought in LLM's perspective is 4, corresponding to "Analyse" category, contrasting with the human observation of level 5 questions "Evaluate".



Plot 1: Comparison of questions classifications made by human agent and GPT-5 with different levels of 'effort' parameter.

3.2 Validate Question Quality



Plot 2: Comparison of issues detection made by human agent and GPT-5 with different levels of ‘effort’ parameter. Notice that “1” represents the identification of a number of significant issues, not necessarily the amount of issues found.

According to Plot 2, the model produced 9 inappropriate significant mistake labels. However, it did not consider “trick-questions” 2, 15 and 17 as flawed, as expected beforehand. Thus, the problem should be on the criteria for meaningful issues.

An interesting example of LLM limitation to abstract information was question Question 4. The question presented a list with a few numbers and an outlier and asked “which measure of central tendency would best represent typical values”. The LLM stated that the question contained 'Ambiguity, oversimplification', but “typical values” was a reasonable way to address “values besides the outlier”, that just demanded basic analysis to realise.

In question 24, the LLM accuses 'Ambiguity, Multiple correct options, Oversimplification'. In fact, the question is about deciding a course of action, given certain information. Given that “Month 1-3 users have 40% churn, Month 4-6 users have 15% churn”, both insights “Users naturally become more loyal over time”, “Early user experience may need improvement” are possible, however, the option that an employee should investigate to reduce churn is the latter, as the first insight would not produce changes to said problem.

Question 28 is another good example of LLM limitation. The program returned "Ambiguity, No correct option" because it didn’t consider the meaning of “typically represent” in the question about the whiskers meaning in a box blot. Their definition is

more elaborated than “min and max values”, as outliers could be the minimum and maximum values. Formally, the answer should be “min and max values inside a tolerance level”, but “typically”, the whiskers are simply “min and max values”.

3.3 Open research question

The method was tested with 2 different level 5 questions in TEXT format, based on the dataset questions 10 and 23: "When would you use a scatter plot instead of a bar chart?", “When would you use a scatter plot instead of a bar chart?".

Three answers with varying quality were produced for each question to each question, the first and second with better quality, and the second with low quality.

Question = "When would you use a scatter plot instead of a bar chart?"

- Answer 1: You would use a scatter plot when you want to see the relationship between two continuous variables. For example, plotting income against education level reveals correlation patterns that a bar chart would hide.
  grade: 10.0
- Answer 2: Scatter plots are useful when the goal is to detect clusters, trends, or outliers in paired data. Unlike bar charts, they show how one variable changes in relation to another.
  grade: 7.5
- Answer 3: A scatter plot is used when you want to count categories. It’s basically the same as a bar chart but with dots instead of bars.
  grade: 1.25

Question = “You're analyzing customer churn and find that Month 1-3 users have 40%churn, Month 4-6 users have 15% churn. What insight should you investigate?”

- Answer1: The high churn in the first three months suggests problems with onboarding or early product adoption. It would be important to investigate user experience and engagement in the first weeks.
  grade: 7.5
- Answer 2: This pattern implies that newer customers are not finding immediate value. You should investigate why early users drop off and what differentiates them from those who stay longer.
  grade: 6.5
- Answer 3: The main insight is that older customers churn more than new ones. Therefore, you should focus only on Month 4–6 users. The main insight is that older

customers churn more than new ones. Therefore, you should focus only on Month 4–6 users.
grade: 0.0

4 CONCLUSION AND FUTURE DEVELOPMENTS

LLM Bloom classifications can offer some insight on the value of human judgement, although seemingly consistent, they were not able to understand the human thought process, to evaluate questions accordingly. Future prompts could explore instructions such as “replicate the human chain of thought to answer to the question”, or even create a “human being” model to solve problem and describe its logical steps to model responsible for classifying, to better identify cognitive processes and Blooms levels.

Bloom framework did not present itself as a good method to assess question quality with LLMs. If on the one hand, they are word-sensitive, on the other hand, the concepts involved in Blooms Levels are not “crystal-clear” and do not show consensus in their explanation across diferent sources [1],[3],[4]. Also, Bloom’s levels lack a database with many examples from each level, that could provide a model with easy and effective learning.

LLMs usually are good at improving level 1 questions to level 2, however, they only change the words without perceiving the thought process of a human being, which leads to errors. Overall, the examples obtained in this study suggest that LLMs are not reliable for the task. However, further study would be necessary to assure that.

LLMs are definitely able to generate interesting question variations, although it was not possible to check LLMs developing optimized questions consistently. LLMs could be good at pointing possible faulty questions, for humans to check the questions that they raised suspicion on, instead of reviewing the whole question database by themselves.

The TEXT correction method a structured way to evaluate free-text answers. It explicitly separates the dimensions of **logical validity** and **contextual relevance**, ensuring that answers are rewarded for being both correct and on-topic. Although simplistic, the approach is interpretable, scalable, and can be improved by refining argument segmentation and weighting schemes.

REFERENCES

1. Institute of Education Sciences. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* [PDF]. U.S. Department of Education. Retrieved from IES website ies.ed.gov

2. OpenAI. (s.d.). *Overview*. In *OpenAI Platform Documentation*. Retrieved August 29, 2025, from OpenAI Platform website: https://platform.openai.com/docs/overview

3. Rutka, J. (2025, August 1). *What Is Bloom's Taxonomy? 100+ question stems & examples*. Top Hat Blog. Retrieved from Top Hat website Top Hat

4. Colorado College. (n.d.). *Bloom's Revised Taxonomy — How to assess learning (learning outcomes)*. Retrieved from Colorado College website