EVALUATION OF AUTOMATIC MULTIPLE CHOICE QUESTION GENERATION USING PROMPT ENGINEERING

Introduction

With the advent of a new technological age over the past few decades, Generative Pre-Trained Transformers (GPTs) have become a centrepiece in online learning due to their autonomy and versatility. This project aims to evaluate and enhance their effectiveness in Automatic Multiple Choice Question Generation (A-MCQG) for maintenance manuals by evaluating the various prompt engineering techniques and GPT models that are readily available for use.

Methodologies

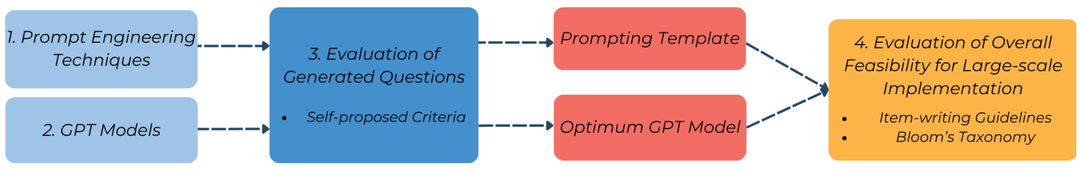

Figure 1: Conceptual Overview of Methodologies

1. Prompt Engineering Techniques: Zero-shot prompt used as baseline. Developed a customised A-MCQG prompting template integrating 4 prompt engineering techniques from Table 1. Systematically removed each technique from template, creating modified prompts for independent testing in Claude.
2. GPT Models (Claude, Bard, and Bing Chat): Prompting template was used in different GPT models to generate a dataset of 20 questions each.
3. Evaluation of Generated Questions: 2 human evaluators identified the number of errors based on 5 parameters: Relevance (25%), Factual Accuracy (30%), Grammar and Readability (10%), Distractor (10%), Answerability (25%).
4. Evaluation of overall feasibility for large-scale A-MCQG implementation: 100 generated questions were evaluated based on 15 frequently occurring Item-Writing Flaws (IWFs) and Bloom's Taxonomy, where questions are classified into two distinct levels of cognition (B1 and B2). B1 assess recall of facts/basic comprehension while B2 assess application/analysis of facts.

Technique	Prompt
Roleplay	You are a creator of highly effective diagnostic quizzes. Your goal is to help instructors create quizzes that can give them a sense of the students' progress.
Context Manager	The quizzes you create will be in a multiple choice format, where each question will have 4 plausible alternatives with no "all of the above" option. Do not make up fake questions or answers and only use information given in the document.
Flipped Interaction	Ask the instructor clarifying questions one at a time to gather details of the quiz such as the sub-topics to be tested. Wait for the instructor to answer before you move on to the next question. Then ask the instructor if they have any other instructions regarding the quiz. Generate the quiz only after asking all the clarifying questions needed to gather information for the quiz.
Fact Check List	When generating the quiz, create a set of facts that can be derived from the attached document. List this set of facts at the end of the output

Table 1: Prompt Engineering Techniques

Results

Evaluation of Prompt Engineering Techniques

Our prompting template showed the highest score of 91.0%. In terms of "Factual Accuracy", "Grammar and readability" and "Distractor", our template showed the highest score of 30.0%, 9.5% and 9.0% respectively. This suggests that it is effective in ensuring coherence and clarity in language while maintaining scientific accuracy in the generated questions. While the "Answerability" of our prompting template (18.8%) was relatively lower than if some techniques were removed, this could be attributed to our template's nature of restricting the replies.

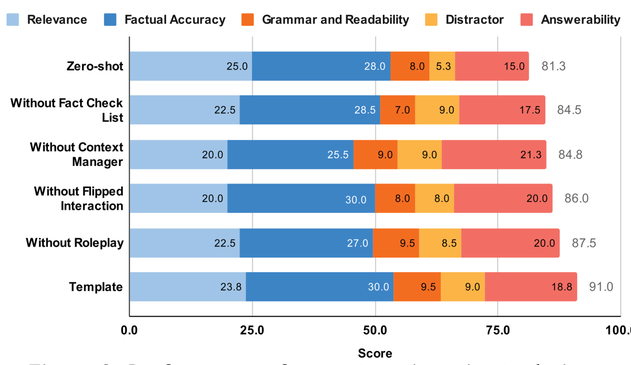

Figure 2: Performance of prompt engineering techniques

Evaluation of Readily Available GPT Models

Questions generated using Claude showed the highest percentage accuracy of 91.0%.

- Relevance to document: Claude showed the score of 23.8%, which suggests that it is most suited for A-MCQG tasks requiring information retrieval from textual documents.
- Answerability: Claude showed the lowest score of 18.8%, which suggests that it is limited in producing plausible distractors. Minor adjustments of the generated distractors can ensure that each question has a single best answer.

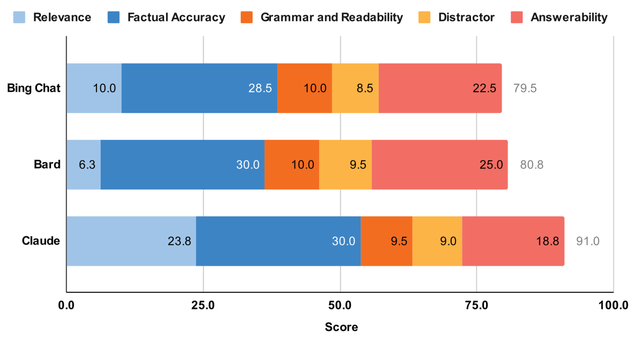

Figure 3: Performance of different GPT models

Evaluation of Overall Effectiveness (Item-Writing Flaws (IWFs))

Number of IWFs	Number detected
None	51
One	27
Two	15
Three	7

Table 2: Total number of IWFs in GPT-generated MCQs

A total of 79 IWFs were detected in the 100 questions generated from our prompt template. Frequency of IWFs (49.0%) is similar to Costello's study of Massive Open Online Courses[1] (MOOCs) (47.4%) and lower than Kirwan's study of MOOCs[2] (54.9%), suggesting that the performance of our prompt template is comparable to traditional methods of generating questions. Furthermore, majority of the generated questions with flaws (27.0%) contained only one IWF, suggesting that the quality of these MCQs could be improved easily with minor edits.

Evaluation of Overall Effectiveness (Bloom's Taxonomy)

Proportion of questions generated at lower cognitive level (B1) and higher cognitive level (B2) were 83.0% and 17.0% respectively. The amount of higher cognitive questions are higher when compared to Mehmood's study pegged at the secondary school level[3] (9.5%) but lower when compared to Momsen's study pegged at the undergraduate level! (37.0%). This can be attributed to (a) nature of content in given document being more informative than application-based and (b) GPTs are limited in their ability to generate questions of higher cognitive domains, which require more creative thinking, leading to challenges in having a standard template across different topics.

Conclusion

The optimal approach for A-MCQG tasks is a combination of roleplay, context manager, fact check list and flipped interaction using Claude. Furthermore, A-MCQG using GPTs is a feasible alternative to manual generation of MCQs and has potential to be implemented into the educational landscape as the quality of GPT-generated questions is comparable to human-generated ones. However, slight modifications to the proposed template and fine-tuning the GPT to specific domains is necessary to generate more relevant and better quality MCQs that can effectively assess learning outcomes.

Limitations

Our use of human evaluation for data collection, combined with our limited number of evaluators, could have led to inter-rater subjectivity in identifying and classifying the errors present. This may have given rise to random errors caused by differing viewpoints, resulting in variability in our result.

Future Work

Future work can explore other forms of question generation like short-answer and open-ended questions, as well as open-source GPT models like Mistral and Phi-2 since they can be hosted on premise and will allow for the generation of questions on classified manuals.

1. Costello, E., Brown, M., & Holland, J. C. (2016). What Questions are MOOCs Asking? – An Evidence-Based Investigation. ResearchGate.
2. Costello, E., Holland, J., & Kirwan, C. (2018). The future of online testing and assessment: question quality in MOOCs. International Journal of Educational Technology in Higher Education, 15(1).
3. Ahmad, M. (2019). Analysis of question papers of Physics at secondary level in Pakistan in the light of revised Bloom's Taxonomy.
4. Momsen, J. L., Offerdahl, E. G., Kryjevskaia, M., Montplaisir, L., Anderson, E. M., & Grosz, N. (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. CBE- Life Sciences Education, 12(2), 239-249.

Members:
Jocelyn Chai Hui Min, Raffles Girls' School
Lu Zhiyue, Raffles Girls' School
Mentor:
How Chang Hong, Defence Science and Technology Agency

YDSP Young Defence Scientists Programme
DSTA Defence Science & Technology Agency
DSO NATIONAL LABORATORIES