Investigating the Use of Intelligent Tutors Based on Large Language Models: Automated generation of Business Process Management questions using the Revised Bloom's Taxonomy

Guilherme Rego Rockembach[1], **Lucineia Heloisa Thom**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{grrockembach,lucineia}@inf.ufrgs.br

Abstract. *The construction of assessment artifacts is a complex task, since generating appropriate assessments manually requires in-depth knowledge of both the area to be assessed and the cognitive processes involved in learning. The use of Large Language Models (LLMs) as the basis for the operation of Intelligent Tutoring Systems can assist in this task. This work experiments with the GPT-3.5-Turbo and LLama-2 LLMs as a source of automatic generation of assessment questions. The experiment was carried out using Prompt Engineering techniques to generate questions for the Business Process Management (BPM) discipline. From the experiment, it was possible to observe that both models are capable of generating questions appropriate to the BPM context. It was also identified that, when it received the context and the model of the question to be generated, the LLama-2 model produced questions more appropriate to the desired cognitive level, while the GPT-3.5-Turbo model received only the context and produced a similar response.*

Resumo. *A construção artefatos avaliativos é uma tarefa complexa, pois gerar avaliações adequadas de forma manual exige um profundo conhecimento, tanto da área a ser avaliada quando dos processos cognitivos envolvidos no aprendizado. A utilização de Large Language Models (LLMs) como base de funcionamento de Sistemas Tutores Inteligentes pode auxiliar nesta tarefa. Este trabalho experimenta os LLMs GPT-3.5-Turbo e LLama-2 como fonte de geração automática de perguntas avaliativas. O experimento foi realizado utilizando técnicas de Engenharia de Prompts na geração de perguntas da disciplina de Business Process Management (BPM). A partir do experimento foi possível observar que ambos os modelos são capazes de gerar perguntas adequadas ao contexto de BPM. Foi identificado também que, quando recebeu o contexto e o modelo da pergunta a ser gerada, o modelo Llama-2 produziu questões mais apropriadas ao nível cognitivo desejado, enquanto que o modelo GPT-3.5-Turbo recebendo apenas o contexto foi possível observar resposta similar.*

1. Introduction

One of the most complex tasks for an educator is assessing student learning. The complexity lies in creating appropriate assessment tools that are capable of effectively measuring the construction of knowledge [Chen et al. 2023b]. Assessment of learning

DOI: 10.5753/sbie.2024.242199

is an ongoing process that encompasses the entire teaching process and is capable of identifying acquired skills and existing weaknesses. Monitoring each student's progress individually is a challenge, especially in distance learning, since contact with the student is limited in this context. Another important factor in the complexity of individual assessment in distance learning is the number of students per teacher, which tends to be much higher than in face to face learning. Tools that assist teachers in the assessment process could minimize this complexity. Intelligent Tutoring Systems (ITS) are an example of a tool that can be applied in this context [Mousavinasab et al. 2021]

ITS are software applications that support learning activities, such as resolving doubts, correcting assignments, and monitoring study progress [Ji and Yuan 2022]. The goal of an ITS is to simulate aspects of a human tutor, often incorporating Artificial Intelligence techniques [Gavidia and de Andrade 2003]. ITS can help in various educational moments and contexts. For instance, ITS can allow students to practice what is being studied in a face to face course, including customized exercises. In the context of Distance Learning and Massive Open Online Courses, ITS can perform learning assessments through Automated Question Generation (AQG).

ITS can also significantly contribute to the education of disciplines such as Business Process Management (BPM) [Dumas et al. 2018]. Challenges in BPM education, as reported in the literature, include the lack of appropriate didactic and pedagogical materials and interactive Information Technology tools to support efficient learning [Moreira et al. 2022, Chow 2021]. Integrating ITS can facilitate the understanding of studied concepts, promoting a more practical education aligned with organizational realities.

A challenge of AQG in ITS is the ability to generate suitable assessment questions, particularly regarding their complexity. The Revised Bloom's Taxonomy (RBT) is often used to measure question complexity [Conklin 2005]. With the recent popularity and advancement in the generalization capability of Large Language Models (LLMs) like GPT, Gemini, and LLama, promising approaches in AQG using these models have emerged [Lee et al. 2023, Pham et al. 2024].

In this context, this paper presents a study that evaluates the capabilities of two emergent LLMs, namely GPT-3.5-turbo, most used LLM chatbots available today [Chen et al. 2023a], from OpenAI [OpenAI 6 20] and LLama-2, one of the most popular opensource LLM models [Sharma et al. 2024], from Meta [Meta 6 20], in the automatic generation of questions according to specific levels of RBT, within the context of BPM education. These LLMs were chosen for this evaluation as they represent two distinct groups of models. GPT is a proprietary LLM, which utilized a large number of parameters in its training. Llama, on the other hand, is an open-source model trained with a reduced number of parameters. This difference makes the GPT model more generic, i.e. capable of performing different tasks, but also increases its dependence on computational resources for execution. This work also assesses the generation capability of less complex models compared to more robust models, by comparing different Prompt Engineering techniques.

This study answers the following Research Questions (RQ):

RQ1. Can LLMs GPT-3.5 and LLama-2 generate good evaluative questions about BPM based on context?

RQ2. Does the use of question templates in the prompt provided to LLM contribute to generating good BPM questions according to RBT?

The remainder of the paper is organized as follows: Section 2 presents the preliminaries, along with reviews of related studies on automated question generation and LLMs. Section 3 describes the methodology used in this study. Section 4 presents the answers for RQ1 and RQ2 and the complementary results. Finally, Section 5 discusses conclusions and potential directions for future work.

2. Preliminaries

In this section, the theoretical foundations that supported this study are presented, as well as works related to AQG and LLMs.

2.1. Theoretical foundation

Although the importance of BPM for organizations is increasing, the search for qualified professionals in this area often proves to be a challenge, even when dealing with professionals with higher education [Silva 2023]. Effective BPM education requires a solid understanding of the methodologies used in the field. Educational strategies based on Information and Communication Technologies can be promising options for facing this challenge.

In this sense, ITS can be used as a tool to support and monitor BPM learning. According to Ji and Yuan (2022) , ITSs are software, typically endowed with Artificial Intelligence technology, that simulate the human tutor acting as an educational mediator. ITSs can provide students with individualized education that meets their specific needs in the learning process. According to the study conducted by Silva et al. (2023), ITSs stimulate self-regulated learning, responsibility, collaboration, teamwork, problem-solving, and motivation [Silva et al. 2023].

ITS can be effective strategies in the context of Distance Learning, since the assessment task in this context can be challenging due to the large number of students that need to be assessed [Xiong and Suen 2018]. In this sense, an ITS equipped with AQG can assess each student's learning through a different set of questions, making the process more meaningful and secure, as students will not know in advance which questions they will receive. Another application of an ITS equipped with AQG is student motivation. In a gamification pedagogical strategy, for example, in which students are rewarded for carrying out extra activities, an intelligent tutor could be implemented to generate these additional tests in an individualized and customized manner, which could contribute to the student's intrinsic motivation [Júnior et al. 2023], both in face to face and distance learning contexts.

According to Mousavinasab et al. (2021), ITSs are generally heuristic-based, that is, dependent on human expertise, which can limit their generalization capacity and consequently hinder the ITS from acting in a personalized manner [Mousavinasab et al. 2021]. This limitation can be mitigated through the use of the great generalization capacity of Generative AI technologies. LLMs are machine learning-based language models with the ability to understand natural language. These probabilistic models of Generative Artificial Intelligence can process and generate text similar to that produced by humans [Filho et al. 2023].

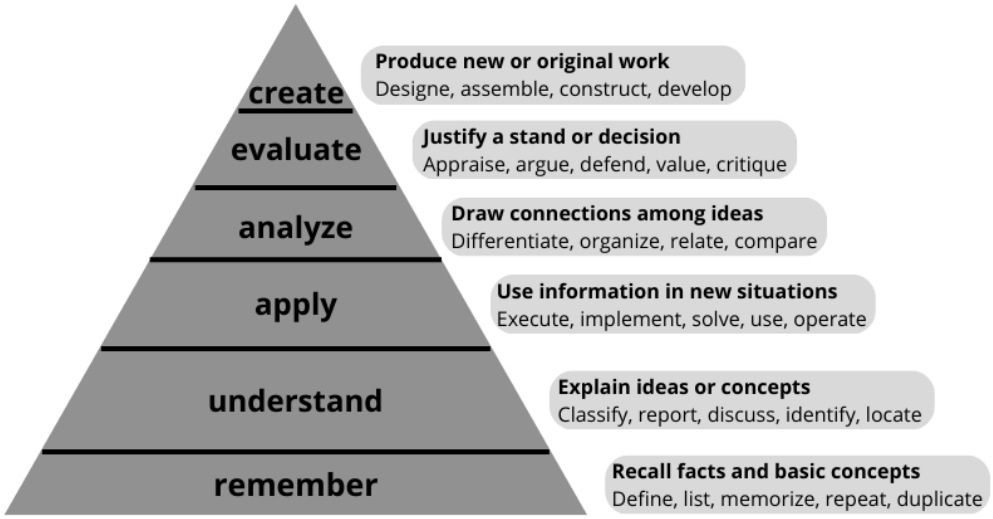

Figure 1. Pyramid with the Cognitive Levels of RBT.

LLMs are classified according to their purpose. Adjusted classified LLMs are used to perform specific tasks, such as translating a text from one language to another. Pre-trained LLMs are those that have a more generic application, that is, they can be applied to a wider range of functionalities, as they have a greater capacity for generalization. Both types of LLMs can suffer from a process called "hallucination", which refers to the model's anomaly in generating meaningless content or content that does not meet the request [Zhang et al. 2023].

A recurring problem when using LLMs to support the functioning of ITSs is that their performance does not remain within the intended pedagogical strategy. According to Chowdhury et al. (2024), this type of ITS can make important errors in generating assessments, such as leaking answers to questions to the student, for example [Chowdhury et al. 2024]. Therefore, it is necessary to ensure that what is generated by the model is aligned with what was planned, preventing "creativity" from hindering its performance. In this sense, the use of Prompt Engineering can mitigate this challenge.

Prompt Engineering techniques involve incorporating the description of the task that the LLM should perform directly into the input, usually in the form of examples of what is expected as output [Gero et al. 2022]. The AQG of an ITS that operates on LLM can be controlled using the Prompt Engineering technique, ensuring that the generated question is in line with its purpose.

One of the important metrics in an assessment question is the cognitive level that the student needs to reach to answer the question, as an assessment must be able to assess the students' diverse skills [Alammary 2021]. The Revised Bloom's Taxonomy [Conklin 2005] can be a suitable parameter in this process since it is widely used in the educational context to guide the evaluative process. In the context of RBT, human cognition can be classified into six hierarchically defined levels on a scale ranging from simplest to most complex, where at each level of the scale, some skills need to be achieved. The levels of RBT can be seen in Figure 1.

In the next section, works related to AQG that have utilized LLMs in different

contexts are presented and discussed.

2.2. Related Works

This section presents a survey of related works on AQG using LLMs. As it is a relatively new field of study, given that the popularization of LLMs is recent, the works related here were published from the year 2022 onwards.

The work conducted by Lee et al. (2023) aimed to improve the quality of English AQGs made by LLMs using Prompt Engineering [Lee et al. 2023]. The study's results indicated that using Prompt Engineering with ChatGPT improved question quality, as validated by specialists and teachers. Feedback from the validation process helped refine the protocol, structure, and question generation prompts, increasing the practicality and accuracy of the system for educational use. Dijkstra (2022) used LLM GPT-3 and Prompt Engineering techniques in creating educational quizzes on reading and text interpretation. Human evaluation of the generated questions revealed that the created quizzes were reasonable, with challenges in generating high-quality distractors (incorrect alternatives) [Dijkstra 2022].

Other works that sought to explore ChatGPT's capability in question generation were those of Pham et al. (2024) and Nasution (2023). In Pham et al. (2024), the aim was the generation of pre-university mathematical questions, while in Nasution (2023), the objective was to evaluate the quality of multiple-choice questions generated by ChatGPT. The methodology employed in both studies was the use of Prompts Engineering, experimenting with the differences between presenting the context to the model or not. The conclusions drawn in the studies highlighted the potential of using LLMs like ChatGPT to generate questions, suggesting that by providing difficulty requirements and demonstrations, ChatGPT could generate better questions [Pham et al. 2024, Nasution 2023].

Previous experiments have been conducted with the aim of fine-tuning LLMs in the generation of subjective questions [Babakhani et al. 2024, Bhat et al. 2022, Sharma et al. 2022]. The LLM used by Babakhani et al. (2024) and Bhat et al. (2022) was GPT-3, while the work of Sharma et al. (2022) used the T5 LLM. The fine-tuning conducted by Babakhani et al. (2024) used sets of text from news media posts and questions generated by humans regarding this news. Among the contributions of the work is the evident complexity of evaluating the generation of subjective questions automatically. The fine-tuning conducted by Bhat et al. (2022) and Sharma et al. (2022) aimed to generate assessment questions based on textual learning materials. Thus, the model's adjustment was conducted through the presentation of educational materials and their respective evaluative questions. Bhat et al. (2022) used as the model's training base content from the Data Science field, Sharma et al. (2022) used content from the computer science field as a whole.

The objective of Maity et al. (2024) was to create a dataset called EduProbe to generate deep and diverse educational questions in the context of school-level subjects. The methodology used was the exploration of prompt-based techniques (no prompt, short prompt, and long prompt) with LLMs to guide question generation. Questions from different areas were generated, including History, Geography, Economics, Environmental Studies, and Sciences. The results obtained in the experiments show that the T5

Table 1. Summary of studies on question generation by LLMs

Study	Goal	Domain	Technique	Results
[Lee et al. 2023]	Improve the quality of AQGs made by LLMs	Teaching English	Prompts Engineering	Improvements in the quality of questions, validated by experts and teachers
[Babakhani et al. 2024]	Get personal preferences based on a text	Opinion Analysis	Fine-Tuning	Improvements to subjective AQGs. Highlights the complexities of evaluating subjective question generation.
[Maity et al. 2024]	Create a dataset called EduProbe to generate deep educational questions	History, Geography, Economics, Environmental Studies and Sciences	Prompts Engineering	LLLMs have shown good performance, but still do not reach human level in most cases.
[Pham et al. 2024]	Analyze ChatGPT's ability to generate pre-university math questions	Mathematics	Prompts Engineering	By providing difficulty requirements and demonstrations, ChatGPT was able to generate aligned questions.
[Nasution 2023]	Evaluate the quality of multiple choice questions generated by ChatGPT	Biology	Prompts Engineering	Human reviews found the AQGs relevant and clear.
[Bhat et al. 2022]	Generate assessment questions based on textual learning materials	Data Science	Fine-Tuning	The GPT-3 model, fine-tuned on the LearningQ dataset, achieved a reasonable level of agreement with expert raters.
[Dijkstra 2022]	Automatic quiz generation in the educational domain	Reading and text interpretation	Prompts Engineering	Human evaluation revealed that the generated questionnaires were reasonable, with challenges in generating high-quality distractors.
[Sharma et al. 2022]	Create question sets with similar difficulty levels for educational assessments in the Distance Education environment	Computer Science	Fine-Tuning	The AQGs generated by the adjusted model automatically evaluated through NLP metrics demonstrate high quality and diversity in the questions created.

model outperforms other models in all automated metrics for question generation [Maity et al. 2024].

Since the study of LLMs is a rapidly growing area, the generation of evaluative questions using LLMs has also received a lot of attention recently, which can be seen in this section. Approaches that use LLMs have the potential to revolutionize the way we learn and teach. However, some challenges in the generation of evaluative questions need to be overcome, especially with regard to the ability to generate quality questions for a wide variety of domains and cognitive levels. Another challenge is the ability to generate automatic evaluators of the generated questions. The studies presented in this section were summarized and can be seen in Table 1. It can be observed that none of the studies presented refer to the automatic generation of questions for BPM, as no examples were found in the literature. This absence highlights the novelty of the study presented here and the need for research on the application of LLMs in AQGs for BPM.

3. Research Methodology

To evaluate the performance of LLMs GPT-3.5 and Llama-2 in automatically generating questions based on context and specific RBT levels, a case study was conducted applying them to generate questions for a specific context within the BPM discipline. This involved introductory BPM content presenting foundational concepts, such as the difference

Create a question, just a question, without an answer, based on the following content:

In BPM, Business Process Management, a business process includes a set of events and activities:

- Events correspond to things that occur atomically, without duration. An event can trigger the execution of a series of activities. Example: inspecting if the received material is correct.
- Task vs. Activity
  - Task: Atomic step in the process. Example: In a purchasing process, verifying if the received equipment matches the specified.
  - Activity: Several related steps, usually consecutive.
  - Automatic Task: Can be automated by a workflow management system.
  - Manual Task: Doesn't support computerized automation.

Use the template below to create the question: {template}

Figure 2. Prompt with context and question template

between tasks and activities, for example.

The objective of the experiment is to answer the following research questions: RQ1: Can LLMs GPT-3.5 and LLama-2 generate good evaluative questions about BPM based on context? RQ2: Does the use of question templates in the prompt provided to LLM contribute to generating good BPM questions according to RBT?

The context from which questions were to be automatically generated was presented to the LLM models using two types of prompts: one presenting only the context and the desired RBT level for question generation (Figure 3), and another providing both the context and a question template related to the desired RBT level (Figure 2). The question models used belong to a set of "Questioning Prompts", referred to in this study as "QP", facilitating their presentation in tables and figures. These models were previously developed by experts at Illinois State University [Illinois State University 6 21] and are aligned with the levels of the RBT.

Five questions were generated for each cognitive level of RBT for each prompt and LLM, totaling 120 questions. Initially, the questions were manually assessed by the authors for their relevance to the requested context. Subsequently, the generated questions were evaluated for their cognitive level alignment using the BloomBERT automatic classifier [Meher and Mall 2023], which utilizes deep learning techniques to classify assessment questions based on their cognitive complexity.

Communication with the GPT-3.5-turbo model utilized the ChatGPT conversation platform provided by OpenAI. Interaction logs with the model are available on the platform for consultation, both for the prompts that presented the context and the

Create questions based on the following content. Use Bloom's taxonomy as a reference to create the questions. Create 5 different questions for each cognitive level in this order:

- remember
- understand
- apply
- analyze
- evaluate
- create

Content:

- Events correspond to things that occur atomically, without duration. An event can trigger the execution of a series of activities. Example: inspecting if the received material is correct.
- Task vs. Activity
  - Task: Atomic step in the process. Example: In a purchasing process, verifying if the received equipment matches the specified.
  - Activity: Several related steps, usually consecutive.
  - Automatic Task: Can be automated by a workflow management system.
  - Manual Task: Doesn't support computerized automation.

Figure 3. Prompt with context only

model of the question to be generated [OpenAI 6 14] and for the prompt that contained only the context [OpenAI 6 13]. Interaction with the Llama-2 model was conducted through the Google Colab platform [Google 6 21a] configured with the NVIDIA L4 GPU. Communication with the model was performed using Python language through the Huggingface repository [Hugging Face 6 21]. Within the Huggingface repository, three different versions of Llama-2 models are available, differing in the number of parameters used for training. For the research objectives, the Llama-2 version trained with 7 billion parameters, called Llama-2-7B, was chosen.

The classification of question levels generated by BloomBERT was carried out through API access on the Google Colab platform. Using the Python programming language, the process was implemented to be accessible to anyone interested [Google 6 21b].

4. Experimental Analysis

The questions generated by LLMs were organized for easy side-by-side comparison in a spreadsheet [Google 6 21c]. In this spreadsheet, it is also possible to view the QPs used in more complete prompts, i.e., those where the desired context and the question model to be generated were provided to the LLM. Each sample was assigned a number to facilitate

Table 2. Percentage of Questions Generated in the Correct Context

Model	Prompt Type	RBT Level	% of questions
GPT-3.5-Turbo	With QP	remember	100
		understand	100
		apply	100
		analyze	100
		evaluate	100
		create	100
	Without QP	remember	100
		understand	100
		apply	100
		analyze	100
		evaluate	100
		create	100
Llama-2-7B	With QP	remember	100
		understand	80
		apply	80
		analyze	60
		evaluate	60
		create	80
	Without QP	remember	100
		understand	100
		apply	100
		analyze	100
		evaluate	100
		create	100

its reference in this paper and ease of consultation by readers.

RQ1. Can LLMs GPT-3.5 and LLama-2 generate good evaluative questions about BPM based on context?

Initially, the questions generated were evaluated by the authors regarding their suitability to the BPM context provided. This analysis was performed by observing each question individually regarding some characteristics: clarity; contextual suitability; structural problems. One of the problems that we sought to identify was the existence of questions that, when generated, brought the answer in the question in the statement, for example. The Table 2 shows the percentage of questions classified as pertinent to the requested BPM context.

The GPT-3.5-Turbo model, as expected due to its greater generalization ability, generated context-appropriate questions for both types of prompts and for all requested complexity levels. However, the LLama-2-7B model had some difficulty generating context-appropriate questions when the question generation prompt contained the question template, at practically all levels of the RBT, with the exception of the least complex level, “remember”, where all questions generated by the model were considered appropriate.

QP: What choice would you have made...?

AQG Llama-2-7B: What choice would you have made when defining the task vs. activity in a business process, given that an event can trigger the execution of a series of activities, and a task is an atomic step in the process, while an activity is several related steps usually consecutive? (sample 23)

AQG GPT-3.5-Turbo: What choice would you have made to ensure tasks are correctly automated in BPM? (sample 83)

Figure 4. Comparison between AQGs GPT-3.5-Turbo and Llama-2-7B from the same prompt

As a smaller-scale language model, that is, one that was trained with less content compared to more robust models, LLama-2-7B may in some cases need more context to generate higher-quality questions, and in these cases, the question models, instead of helping, act as a limiting factor. In one of the samples, in order to meet the requirements for generating the question according to the requested model, LLama-2-7B presented a "hallucination" and included its answer in the question itself, as reported by Chowdhury et al. (2024). In Figure 4 the occurrence of the "hallucination" can be observed.

As observed in this experiment, both tested models demonstrate the ability to generate questions effectively within the given context. However, the utilization of less complex LLMs necessitates additional attention, particularly when employing Question Prompts. It becomes essential to provide sufficient context or select a Question Prompt that aligns more appropriately with the given context to enhance the quality and relevance of the generated questions. Fine-tuning these parameters can significantly optimize the performance and usability of automated question generation systems in educational and practical applications.

RQ2. Does the use of question templates in the prompt provided to LLM contribute to generating good BPM questions according to RBT?

To answer RQ2, we classified the AQGs using the BloomBERT automatic classifier. This model was trained with over 110 thousand questions labeled by RBT, achieving an accuracy of approximately 75%. The aim of this classification was to compare the AQGs generated from two different types of prompts: those that included only the context and those that, in addition to the context, presented a specific question model, referred to as "Questioning Prompts", formulated by education professionals to be within the expected cognitive level.

The adherence index of each approach to the desired RBT cognitive level can be observed in Table 3. To present the different types of prompts used in Table 3, the acronyms NPQ were used to identify prompts that presented only the context and WPQ for prompts that, in addition to the context, presented a model of the question to be generated.

From Table 2, we can observe that including question models in the prompt for GPT does not seem to increase the adequacy of questions to the desired cognitive level.

Table 3. Adherence Index to RBT

Model	Prompt	Rem.	Und.	App.	Ana.	Eva.	Cre.	Total
GPT-3.5-Turbo	NPQ	40%	80%	20%	80%	80%	100%	66%
	WPQ	40%	60%	20%	20%	40%	80%	43%
LLama-2-7B	NPQ	40%	60%	20%	20%	0%	60%	33%
	WPQ	20%	20%	60%	60%	20%	80%	43%

With QP What could you invent to better automate manual tasks in BPM? (sample 90)

Without QP: Formulate a proposal for enhancing the flexibility of BPM systems by incorporating real-time event monitoring and response capabilities. (sample 120)

Figure 5. Comparison between AQGs GPT-3.5-Turbo Create level of RBT

In all RBT levels, GPT performed better when it received only the context and desired cognitive level. However, although GPT-3.5-Turbo is capable of generating appropriate questions in terms of cognitive level, the use of Questioning Prompts can help control the difficulty level of questions.

In Figure 5, we can compare two questions generated by GPT-3.5-Turbo for the "Create" cognitive level: one using a question model and the other only with the context and desired level. It is evident that the model generated a more complex question when it did not use a question model. In this sample, it is evident that when GPT-3.5-Turbo has complete freedom to create questions for a given context and cognitive level, it can generate questions that, while matching the desired Revised Bloom's Taxonomy level, may be more complex than questions traditionally created by humans, which can be a problem.

It was observed that, in some cases, GPT-3.5-Turbo generated questions using only the context that were very similar to those generated using question models. This can be seen in samples 68 (with QP) and 92 (without QP), where the main verb of the question is "differentiate," asking the respondent to differentiate tasks and activities. According to the experts who created the QPs, questions with this purpose, i.e., differentiation, belong to the "understand" level of Revised Bloom's taxonomy. However, GPT-3.5-Turbo generated sample 92 as being at the "remember" level.

For LLM Llama-2-7B, the results in Table 3 show that question models prompts were particularly effective in generating appropriate questions for higher complexity levels of RBT. On average, using prompts with question models is preferable to using them without. Furthermore, considering that the classifier makes classification errors, since its accuracy is not 100%, we can perform a relaxed analysis regarding its classification. If we consider as correctly classified the questions that were categorized with one of the levels subsequent to those that they actually correspond to, the samples that had worse results improve significantly, as in the case of the questions at the "evaluate" level, which go from 20% to 80% of the questions generated correctly.

Similar to GPT-3.5-Turbo, in some samples LLama-2-7B also generated questions using the verb "differentiate" for the cognitive level "understand" of the RBT. However, LLama-2-7B also generated questions using the same verb for the cognitive level "remember", both samples in cases where the model received only the context via prompt (samples 34 and 35). This behavior of the model highlights an inconsistency in the way it determines the characteristics of the questions for each level, accent the importance of using question models when working with LLMs of this type.

The results of this experiment highlight the potential of using question models to generate questions for less complex LLMs. Using this technique, simpler models are able to generate questions that are appropriate to the BPM context and according to the desired complexity. This result demonstrates the viability of a more economical approach, which can be crucial in educational environments, given that they are generally spaces with few resources.

5. Conclusions and Future Work

Through this investigation, it was possible to conclude that the tested LLMs have the ability to generate appropriate evaluative questions based on BPM contexts. It is also concluded that GPT-3.5-Turbo does not require question models to produce contextually and cognitively appropriate AQGs, which can be extended to other LLMs that have similar characteristics.

Another important finding is that the use of question templates helped Llama-2-7B generate questions that were more aligned with the desired cognitive levels. This conclusion can be extended to LLMs that, like it, were trained with a smaller number of parameters. Thus, it is clear that the selection of Questioning Prompts and appropriate contexts can directly influence the quality of the question generated. In this context, we can conclude that the use of prompts that have, in addition to the context, the model of the question to be generated has the potential to improve the questions generated by LLMs. Thus, investing in ITS approaches that use these techniques may be a possibility to extract better results from less complex LLMs in contexts with limited hardware.

As future work, we intend to develop a prototype of an ITS that uses LLMs of lesser complexity. The purpose is to explore how we can adapt the use of LLMs so that they can meet the specific demand of generating educational questions, in an accessible and scalable platform. To achieve this goal, we also intend to conduct studies that evaluate the quality of the questions generated from different metrics and evaluators, both automatic and human. The current work was important to identify the strengths and weaknesses of the use of LLM engines for ITSs, enabling future adjustments.

Acknowledgment

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Alammary, A. S. (2021). Losmonitor: A machine learning tool for analyzing and monitoring cognitive levels of assessment questions. *IEEE Transactions on Learning Technologies*, 14(5):640–652.

Babakhani, P., Lommatzsch, A., Brodt, T., Sacker, D., Sivrikaya, F., and Albayrak, S. (2024). Opinerium: Subjective question generation using large language models. *IEEE Access*, 12:66085–66099.

Bhat, S., Nguyen, H., Moore, S., Stamper, J., Sakr, M., and Nyberg, E. (2022). Towards Automated Generation and Evaluation of Questions in Educational Domains. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 701–704. International Educational Data Mining Society.

Chen, Y., Arunasalam, A., and Celik, Z. B. (2023a). Can large language models provide security & privacy advice? measuring the ability of llms to refute misconceptions. In *Proceedings of the 39th Annual Computer Security Applications Conference*, pages 366–378.

Chen, Z. et al. (2023b). Student performance prediction approach based on educational data mining. *IEEE Access*, 11:131260–131272.

Chow, W. (2021). Teaching business process management with a flipped-classroom and problem-based learning approach with the use of apromore and other bpm software in graduate information systems courses. In *2021 IEEE International Conference on Engineering, Technology Education (TALE)*, pages 1–8.

Chowdhury, S. P., Zouhar, V., and Sachan, M. (2024). Scaling the authoring of autotutors with large language models. *arXiv preprint arXiv:2402.09216*.

Conklin, J. (2005). Review of *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives Complete Edition*. *Educational Horizons*, 83(3):154–159.

Dijkstra, R. e. a. (2022). Reading comprehension quiz generation using generative pre-trained transformers. In *iTextbooks@ AIED*, pages 4–17.

Dumas, M. et al. (2018). *Fundamentals of Business Process Management*. Springer-Verlag.

Filho, L. P., Souza, T., and Paula, L. (2023). Análise das respostas do chatgpt em relação ao conteúdo de programação para iniciantes. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1738–1748, Porto Alegre, RS, Brasil. SBC.

Gavidia, J. J. Z. and de Andrade, L. C. V. (2003). *Sistemas tutores inteligentes*.

Gero, K. I., Liu, V., and Chilton, L. (2022). Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 1002–1019. ACM.

Google (Accessed: 2024-06-21a). Google Colaboratory Shared Notebook. `https://colab.research.google.com/drive/1sXPqR-0Yycm6li43Urnxru8JMgQz4a5K?usp=sharing`.

Google (Accessed: 2024-06-21b). Google Colaboratory Shared Notebook. `https://colab.research.google.com/drive/1DeeYslSS5ZD2U2cFjdCuYE8QNZrwIAjO?usp=sharing`.

Google (Accessed: 2024-06-21c). Google Sheets. `https://docs.google.com/spreadsheets/d/`

1c-fN01AoxfIwAiiaQfFgoVbrp6jUvDrfQPTVr8EOo8Q/edit?usp=sharing.

Hugging Face (Accessed: 2024-06-21). Llama 2 7B HF Model on Hugging Face. https://huggingface.co/meta-llama/Llama-2-7b-hf.

Illinois State University (Accessed: 2024-06-21). Revised Bloom's Taxonomy. https://education.illinoisstate.edu/downloads/casei/5-02-Revised%20Blooms.pdf.

Ji, S. and Yuan, T. (2022). Conversational intelligent tutoring systems for online learning: What do students and tutors say? In *2022 IEEE Global Engineering Education Conference (EDUCON)*, pages 292–298. IEEE.

Júnior, C. P., Santos, H., Rodrigues, L., and Costa, N. (2023). Investigating the effectiveness of personalized gamification in enhancing student intrinsic motivation: an experimental study in real context. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 838–850, Porto Alegre, RS, Brasil. SBC.

Lee, U., Jung, H., and Jeon, Y. e. a. (2023). Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. *Education and Information Technologies*.

Maity, S., Deroy, A., and Sarkar, S. (2024). Harnessing the power of prompt-based techniques for generating school-level questions using large language models. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 30–39, New York, NY, USA. Association for Computing Machinery.

Meher, J. P. and Mall, R. (2023). Bloombert: A deep learning-based cognitive complexity classifier of assessment questions. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 318–322.

Meta (Accessed: 2024-06-20). LLAMA 2. https://llama.meta.com/llama2/.

Moreira, S. A. S., Sousa, R. G., and Pádua, S. I. D. (2022). DimensÕes para o ensino de business process management (bpm): proposta de um modelo conceitual qualitativo. In *XXV SEMEAD - Anais*, São Paulo. SemeAd.

Mousavinasab, E., Zarifsanaiey, N., Rakhshan, M., Mirzaee, M., Amini, M., and Ghazi Saeedi, M. (2021). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1):142–163.

Nasution, N. E. A. (2023). Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education*, 2(1).

OpenAI (Accessed: 2024-06-13). ChatGPT Shared Link. https://chatgpt.com/share/2f5041fe-1463-480b-9bed-bb5344d9c748.

OpenAI (Accessed: 2024-06-14). ChatGPT Shared Link. https://chatgpt.com/share/db1f8017-c7b3-43e3-8cdc-cb216815cd54.

OpenAI (Accessed: 2024-06-20). OpenAI GPT-3.5 Turbo Documentation. https://platform.openai.com/docs/models/gpt-3-5-turbo.

Pham, P. V. L., Duc, A. V., Hoang, N. M., Do, X. L., and Luu, A. T. (2024). Chatgpt as a math questioner? evaluating chatgpt on generating pre-university math questions. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, page 65–73, New York, NY, USA. Association for Computing Machinery.

Sharma, R. K., Gupta, V., and Grossman, D. (2024). Spml: A dsl for defending language models against prompt attacks. *arXiv preprint arXiv:2402.11755*.

Sharma, S., Agarwal, R., and Mittal, A. (2022). Generating educational questions with similar difficulty level.

Silva, C., Moreira, T., Fernandes, I., Passos, C., Duarte, J., and Goldschmidt, R. (2023). Sistemas tutores inteligentes na aprendizagem por competências: Uma revisão sistemática da literatura. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1120–1132, Porto Alegre, RS, Brasil. SBC.

Silva, D. (2023). Metodologias e abordagens para o ensino e aprendizado de gerenciamento de processos de negócio: uma revisão sistemática da literatura. Dissertação de mestrado, Universidade Federal do Rio Grande do Sul, Instituto de Informática, Porto Alegre. Disponível em: `https://lume.ufrgs.br/handle/10183/263302`.

Xiong, Y. and Suen, H. K. (2018). Assessment approaches in massive open online courses: Possibilities, challenges and future directions. *International Review of Education*, 64(2):241–263.

Zhang, Y. et al. (2023). Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.