Evaluation of LLM Generated Multiple Choice Questions Motivated by Bloom's Taxonomy.

Megan Morris

Master of Science
School of Informatics
University of Edinburgh
2024

Abstract

Developing questions for exams and classwork is currently a time consuming task for teachers and instructors [35], however there is potential for large language models to substantially reduce the amount of time educators must spend developing content for their courses. This research explores the potential of large language models (LLMs), specifically GPT-4o, to automate the generation of multiple-choice questions (MCQs) aligned with Bloom's Taxonomy and assesses their quality for classroom use. The study involved adapting a Python-based application that manages the prompting and parsing of responses from GPT-4o, requiring a specified Bloom's Taxonomy level and a textbook section as inputs for generating MCQs. The generated questions were evaluated for their alignment with Bloom's levels through both human prediction by educators and through a mix of supervised machine learning and deep learning models . Results indicated that while the LLM-generated MCQs were rated highly by educators in terms of quality, they did not consistently align with the intended Bloom's levels, highlighting a significant challenge in this area. We propose that this automated system of question generation could be used as a mechanism for educators to reduce their workload by quickly and efficiently developing high quality questions for their courses.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 643533
Date when approval was obtained: 2024-06-06

The combined participants' information sheet and consent form are included in the appendix.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Megan Morris*)

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Kobi Gal and to Dr. Avi Segal for their mentorship and support throughout my time working on this project. I would also like to thank Tal Yifat at ChalkTalk who created the original version of EduGenie that made this project possible. I would also like to acknowledge the secondary school educators across Scotland and to the NB team at UC Davis who gave their time to evaluate our generated questions. Their insight has been invaluable in shaping this project. Lastly, I would like to thank my friends and family for their love and support throughout my educational journey.

Table of Contents

Chapter 1

Introduction

The role of teachers and lecturers far extends beyond the time spent speaking in front of a class. Leading a course requires a significant time investment in generating teaching material such as course syllabi, lesson plans, and learning objectives for the class [39]. A study conducted on teachers in Germany [35] found that teachers on average spent over 17 hours a week on lesson planning, preparing, and correcting tests and homework—tasks that do not include interacting with students. This substantial investment of time underscores the importance of finding efficient ways to reduce the workload associated with these activities.

The recent development and increased accessibility of large language models offer opportunities to ”significantly ease the workload of educators” by automating much of this content creation, thereby freeing teachers to focus more on direct instruction and student engagement [37]. These models can assist in generating various teaching materials, including lesson plans and test questions, which can be customized to meet specific educational needs.

One particularly promising application is the automatic generation of multiple-choice questions (MCQs), which offers a potential means to simplify a difficult and time-consuming task: creating challenging test questions. MCQs are an attractive method for student evaluation as they are a reliable form of assessment, can be easier to administer and grade than open-ended questions, and can effectively measure a range of cognitive skills [26]. However, creating high-quality MCQs requires significant effort and expertise [13]. Automating this process could greatly benefit educators, provided that the generated questions are of sufficient quality. The flexibility of large language models makes them an ideal tool for approaching this task, as they can be trained to generate questions that are not only standardized for a class at large but also

personalized for individual students based on their current level of understanding and any misconceptions they may have about a subject [38].

This dissertation seeks to investigate the capacity of Large Language Models to generate high quality MCQs that could be used within an educational setting. The generation of these questions will be informed by Bloom's Taxonomy, a framework of six cognitive levels, with each level increasing in its cognitive complexity [14]. The taxonomy consists of six levels- Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating- ranging from basic recall to complex synthesis of information. It is widely used in education to design curricula and assessments that promote increasingly sophisticated cognitive abilities [40]. Challenging questions require students to make connections across topics, analyse situations and to go beyond memorization of facts [7]. The inclusion of different cognitive levels as outlined by Bloom's taxonomy will be used to instruct the LLM to generate questions that are semantically challenging and that require student's to interact more deeply with the content in order to find the correct response.

This dissertation will be investigating the following three research questions.

RQ1: *"How can the process of creating multiple choice questions be automated to support lecturers in teaching?"*

RQ2: *"Can LLMs generate multiple choice questions that correspond to specified Bloom's Taxonomy levels?"*

RQ3: *"Are multiple choice questions generated by large language models of a high enough quality to be usable within a classroom setting?"*

The structure of this dissertation is as follows: Chapter 2 provides an overview of previous research conducted on the use of LLMs in an educational context and on question generation using LLMs. Chapter 3 details the processes used within this project for generating and evaluating multiple choice questions. Chapter 4 will outline and examine the results achieved with focus on the quality level of the generated MCQs and the accuracy with which the LLM generated questions of a desired Bloom's level. Chapter 5 describes the limitations present within this project and contemplates opportunities for future improvement of the work conducted.

Chapter 2

Background

This background section explores the application of large language models (LLMs) in education, focusing on their role in automating content creation, enhancing personalized learning, and generating multiple-choice questions (MCQs) aligned with Bloom's Taxonomy. It also examines research on evaluating the quality of LLM-generated educational content, highlighting both automated and human assessment methods to ensure the effectiveness of these tools in instructional settings.

2.1 Large Language Models in Education

The application of large language models (LLMs) in education has proven to be an effective approach for automating the generation of teaching materials, significantly reducing the time teachers spend on content creation while enabling personalized learning experiences for each student.

Although LLMs are still relatively new in the educational sector, they have been successfully employed to assist teachers in drafting course materials [3], labeling assessment questions [2], and creating standardized assessments [39]. Remarkably, LLMs have demonstrated the ability to generate complete course content up to 25 times faster than traditional methods, with performance that is comparable to human-created content [20]. These applications offer substantial time-saving benefits for educators.

Furthermore, numerous researchers have explored the potential of LLMs to deliver personalized learning experiences, particularly through the development of individualized intelligent tutoring systems [21][3][30]. For example, in [8], the authors focus on creating a tutoring system capable of assessing various cognitive abilities in students, and knowledge of these abilities is then used to tailor the tutoring model to better meet

each student's needs.

While there have been many successful implementations of LLMs into the educational field, there are several potential pitfalls that researchers have noted with regards to generating educational content such as the lack of pedagogical knowledge [30], a tendency towards standardization of material [37] and the risk of hallucinations [46]. While LLMs may be trained on extensive domain-specific knowledge, they have not necessarily been trained with the pedagogical understanding necessary for high-quality educational materials as models trained only on specialist knowledge may produce content that lacks the instructional understanding required for effective teaching [30]. One way to address this challenge may include integrating teaching strategies into prompt design [5] and refining training data to include pedagogical principles [11].

Automatically generating content often relies on using similar prompting techniques repeatedly, and some researchers fear that instead of moving towards personalizing content for individual students, using LLMs for educational content may result in a "standardization that might stifle creativity and innovation in teaching methods" [37]. This concern highlights the need to balance using LLMs for efficiency while allowing for human input to preserve creativity in content development. One way to balance human input with automatic generation, which I explore in this dissertation is to use feedback provided by educators on the content generated by the LLM to refine the content to the needs of the instructor.

Another concerning issue with the usage of LLMs in education is the risk of hallucinations, which occurs when LLMs generate output that seem plausible but that deviate either from the context given or from factual knowledge [46] [30]. In an educational setting, hallucinations can be particularly dangerous, as students encountering new information for the first time may accept incorrect information presented by an LLM as fact, leading to a fundamental misunderstanding of the material. Proposals for how hallucinations in LLMs can be addressed include enhancing the LLMs perception of its capability boundaries so that the model does not generate answers beyond the information it has been trained on [42], adding explanations for the content it has created [17] and injecting knowledge into the prompt to prevent hallucination [3].

Any system developed to generate high quality educational content using LLMs must give careful attention to mitigating hallucinations and informing the model with pedagogical expertise. By ensuring prompts are enriched with relevant knowledge and contexts, developers can create better tools to support effective teaching.

2.2 Bloom's Taxonomy

Bloom's taxonomy is a hierarchical classification of cognitive skills that educators widely use to guide the creation of questions and learning objectives [22][44]. The taxonomy comprises six levels of cognitive skills: remembering, understanding, applying, analyzing, evaluating, and creating. Each successive level in the taxonomy demands a deeper understanding and higher cognitive complexity, making the taxonomy a valuable framework for structuring educational assessments and instructional strategies [29]. Educators often utilize Bloom's taxonomy not only to design challenging questions but also to ensure that students engage in higher-order thinking, promoting critical analysis and creative problem-solving skills. The separation of levels helps to differentiate between learning objectives and questions that only depend on "factual recall" from those that "probe higher-order thinking skills" in order to support a deeper conceptual understanding of course material [44].

Previous research on generating educational content with LLMs has focused on leveraging Bloom's taxonomy to enhance the complexity of generated content, with the goal of progressively developing students' cognitive abilities [16][5][11]. These studies have investigated the potential of LLMs to generate content aligned with various levels of the taxonomy, thereby facilitating the creation of more sophisticated and effective educational tools. Some studies have injected information about Bloom's taxonomy into their prompting strategies without specifying the specific level a question should target [5] while others have also included within their prompt the Bloom's level that the question should correspond to [16]. This dissertation will take the second approach whereby the desired Bloom's taxonomy level acts as one of the inputs into the LLM. In this manner, not only will questions created be informed by Bloom's taxonomy, but users of the system will also be able to specify whether generated questions should be less or more cognitively complex depending on the Bloom's level specified. In doing so, this approach allows will allow us to test not only the quality of the questions generated but also whether LLM-generated questions correspond with the inputted Bloom's level.

2.3 Automatic Question Generation

Prior to the development of LLMs and deep learning approaches which allow for flexibility in their generation, research on question item generation typically relied on identifying patterns within questions and forming a template for the question generation

process [1]. Since LLMs have become more accessible there has also been a significant research focus on how to train LLMs to answer given questions, as opposed to generating the questions themselves [37].

As opposed to open response questions, multiple choice questions pose an interesting challenging for LLMs as the model must be able to individually generate each component of a question well while maintaining overall coherence. There are three main components to a multiple choice question. The stem is the question itself, the key is the correct answer and the distractors are the incorrect answer options [18]. Generating high quality distractors is a problem unique to multiple choice questions. Quality distractors play a role in controlling how difficult a question is and they allow for a question to discriminate between the different misunderstandings about a concept that students may have [18]. In order to be effective, the distractor options must be ”plausible enough to mislead students” but they must not be so ”evidently incorrect as to be easily discernible” [26]. To ensure that distractors are difficult enough to discern so that answering the question requires understanding of the tested concept, distractor options should be ”semantically related” to the key [45]. For example, the wrong options should not be significantly shorter or longer than the correct answer and the syntactic structures between the key and distractors should be similar. As there has been a noted tendency for LLMs to generate low quality distractors, by for example listing correct answers as distractors [38], some researchers have focused specifically on generating distractors given a question stem and correct answers as opposed to generating full multiple choice questions [26] [4] [41].

Previous studies have shown success in generating multiple choice questions using LLMs, with LLM generated questions even being adopted into popular standardised exams [7]. Tran found that GPT-4 produced higher quality multiple choice questions and corresponding answer choices than GPT-3 [38]. Studies comparing LLM generated MCQs against those created by humans have shown that LLM generated questions can be of a comparable quality to human generated questions [10] [31]. Elkins determined found that the majority of questions their study generated with LLMs were rated by educators as either useful with minor edits or useful with no edits [11]. To increase the effectiveness of LLM generated MCQs, prompt engineering is essential to clarify the requirements for a high quality question. Injecting knowledge into the prompt by adding textbook sections for context information [31], instructing the model on pedagogical techniques such as Bloom’s taxonomy [5] and by providing information on what makes a high quality distractor [26] can help improve the quality of questions and

answers generated by the model. In this research project, the prompting strategy will be injected with knowledge about how to generate effective multiple choice questions and information about each of the specified Bloom's levels. In addition, this research will also explore opportunities for integrating feedback from educators as an additional means of knowledge injection.

2.4 Evaluating Generated Questions

The evaluation of educational questions generated by deep learning models prior to the widespread adoption of LLMs predominantly relied on metrics that assessed the syntactic and lexical similarity between the generated output and a human-created reference texts[26]. These traditional NLP evaluation metrics, such as BLEU, ROUGE, and METEOR, emphasize the closeness of the generated response to a pre-defined human response. [23]. However, these metrics pose challenges when applied to LLMs, which generate more flexible and diverse responses. Such models can produce correct and valuable content that may still score poorly on traditional metrics due to differences in structure or phrasing compared to the reference text. Therefore, it's essential to adopt more nuanced evaluation methods that consider the utility and correctness of the generated material rather than just its similarity to human examples.

To address this, automatic evaluation methods have been developed where the LLM-generated multiple-choice questions (MCQs) are evaluated by either the same LLM or a different one. [5]. This automated assessment can identify and filter out low-quality questions based on criteria such as sentence complexity, clarity, and the distinctness of answer choices. [27]. Additionally, human expert evaluation remains crucial. Teachers and instructors assess the quality of questions by considering whether they are informative, accurate, and provide clearly distinct answer choices [10] [31]. This human judgment approach ensures that the generated questions meet educational standards, and it will be employed in this project to ensure the validity and utility of the questions generated by LLMs.

Chapter 3

Methodology

This chapter outlines the methodology employed in this project for automatic question generation. It begins by detailing the process used to generate multiple-choice questions (MCQs) with large language models (LLMs). Next, it describes the dataset of LLM-generated questions that was curated for subsequent evaluation. Finally, the chapter documents the methods used to assess both the quality of the generated questions and their alignment with Bloom's Taxonomy levels, utilizing both human evaluators and supervised machine learning models.

3.1 Question Generation

An application named EduGenie, initially developed in [5] was adapted for the purposes of this research. This application, built in Python using LangChain, generates MCQs by accessing the GPT-4o LLM through OpenAI's API. Additional features that allow for feedback integration, for inclusiong of example questions and for automatic saving of questions were developed as part of this project.

The generation model relies on the user to input the desired number of questions to generate, a specified Bloom's taxonomy level and the textbook section that will be used to generate the questions. Within an "Educational Context" section of the application, the user can also optionally input example questions and a learning goal that the generated questions should address.

Built into the application are instructions about how to generate high quality MCQs and information about how to develop questions at different Bloom's levels. This information is then combined dynamically with user input into one prompt involving a system prompt and a human prompt that is sent to GPT-4o through the API. The output

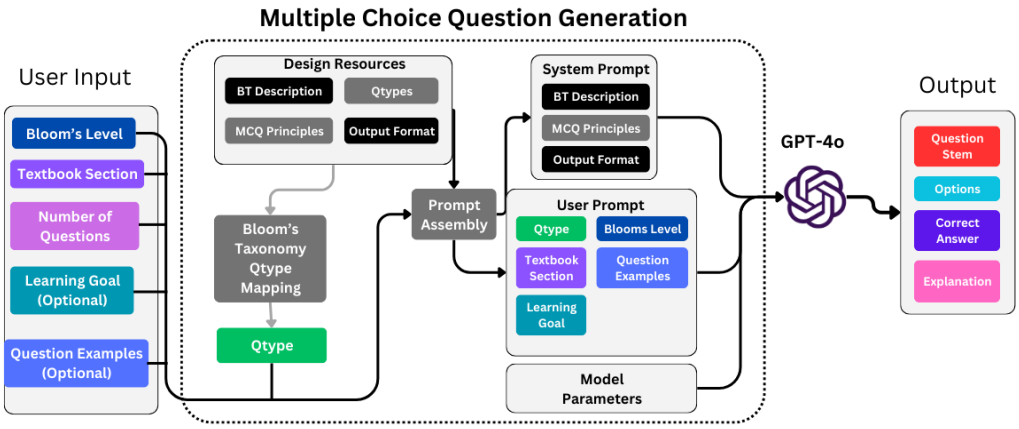

Figure 3.1: This diagram, based on [10], describes the automatic question generation process that was used to generate all of the multiple choice questions examined within this project. For each generation, the user must provide the desired number of questions, a specified Bloom's taxonomy level and the textbook section that the questions will be based on. The user also has the option to include learning goals and examples of other MCQs. The user input and the design resources saved in the system are combined into a system and user prompt submitted to the GPT-4o model which then outputs the generated MCQ(s).

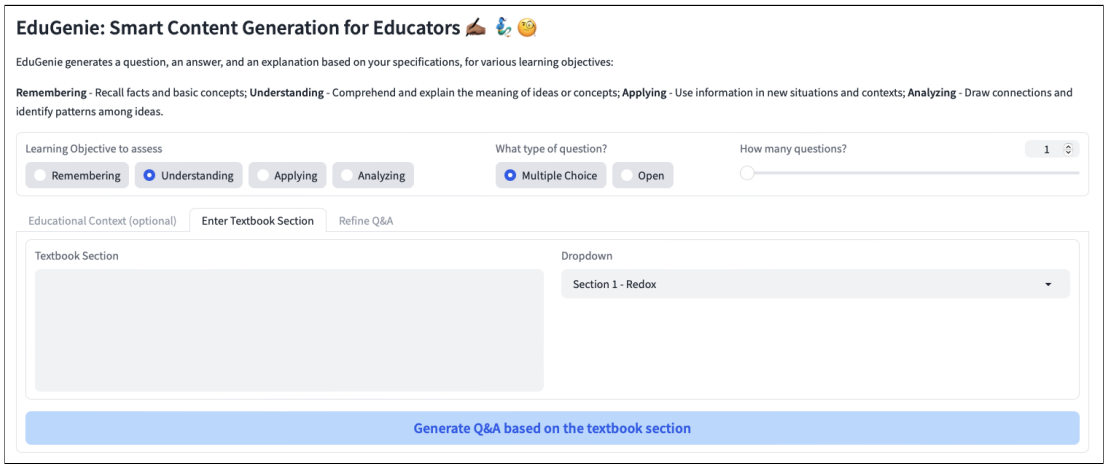

Figure 3.2: The user interface for generating questions with the EduGenie tool.

Human Template

Develop {N} question/s based on the requirements below and the textbook section delimited by triple backticks. Note: learners may not have access to the textbook section, so avoid making references to it.
{educational_context}
{examples}
{q_type}
{learning_objective}
{format_instructions}
Textbook Section:
‘‘‘
{textbook_section}
‘‘‘

Figure 3.3: The human template is generated using a combination of user input and system resources that are selected for inclusion based on the question options chosen by the user.

generated by the model includes a question stem, a list of answer options, the correct option and an explanation as to why the correct option was selected. As can be seen in figure 3.3, the human template relies on a number of input variables as described below.

- **N:** The number of questions to generate.
- **educational_context:** When the user inputs a learning goal, the following phrase is included as the educational context: ”The question should assess whether the student has mastered the specified learning goal. Goal: {learning_goal}”, where the learning_goal variable is the learning goal inputted by the user. If no learning goal is specified, educational_context is set to an empty string.
- **examples:** When the user inputs optional examples of questions, the following phrase is inserted as the examples variable: ”The following are human generated questions optimized for student learning. The generated examples should be of a similar level of difficulty and have a similar, but not identical style to the example questions.: {question_examples}”, where the question_examples variable is the example questions included by the user. If no examples are given, the examples

variable in the human template is set to an empty string.

- **q_type:** The user has the option of having the system generate multiple choice questions or open response questions, however only multiple choice questions are considered. The q_type variable for multiple choice questions provides a detailed description of how to create a high quality multiple choice question. This includes providing instructions to create a focused stem, use plausible distractors and to avoid grammatical clues that give away the correct answer. The full prompt included for multiple choice questions with this variable can be found in appendix A.1.

- **learning_objective:** The user must specify the Bloom's taxonomy level that the generated question should correspond to. Depending on the level specified, the system has different prompts which explain how to generate a question at each of the levels. For example, the prompt for the Applying level is the only level that instructs the model to create a scenario upon which an application question should be based. The prompts for understanding and analyzing provide different strategies that the model should use when creating the question. The Bloom's level selected by the user determines the prompt which is inserted into the learning_objective variable from among the prompts listed in appendix A.2.

- **format_instructions:** The format_instructions variable provides a description of the JSON format that the question should be outputted in with a detailed breakdown of the components that should be included in the output (question stem, options, correct answer and explanation).

- **textbook_section:** The complete textbook section as inputted by the user.

While some studies have prompted large language models (LLMs) with only the desired learning objectives or general topics that the questions should be based on [10], the approach in this project differs significantly. Our approach relies on injecting knowledge into the model through detailed explanations on generating multiple-choice questions and matching them to corresponding Bloom's levels, thus teaching the model about pedagogical principles. Additionally, domain knowledge is provided in the form of the textbook section [30].

The system template in figure 3.4 and the human template in figure 3.3 are concatenated into one prompt that is submitted to GPT-4o using a default temperature setting of 0.7. As the temperature of LLM's range from 0 to 1 with 0 resulting in the most

deterministic outputs, 0.7 was selected to allow for greater creativity within generated responses [30]. The generated questions were then saved into a dataset for evaluation.

The application was also modified to allow questions to be revised based on given feedback. To revise a question, the user needed to input the question, its options, the correct answer and explanation along with the textbook section it came from. The user could also optionally input any learning goals the question is meant to address. The user inputs a rating between 1 to 5 for the question with 1 being the worst and 5 being the best and any feedback on the question that should be used to refine the question. The feedback prompt was then concatenated with the system template into one prompt submitted to GPT-4o with the same default temperature setting of 0.7. The full template that includes user inputted feedback can be found in appendix A.3.1.

3.2 Datasets

Five textbook sections were selected from an open source university-level introductory biology textbook to use for question generation [12]. We generate questions only according to the first 4 Bloom's taxonomy levels: remembering, understanding, applying and analyzing.

The questions generated for the first four sections were used for a first round of evaluation, in a survey given to secondary school science teachers.

Textbook Section	**# Questions**	**Remember**	**Understand**	**Apply**	**Analyze**	**With Learning Goals**
Section 1	16	4	4	4	4	4
Section 2	16	4	4	4	4	4
Section 3	16	4	4	4	4	4
Section 4	16	4	4	4	4	4
Section 5	16	4	4	4	4	4

Table 3.1: The number of questions generated from each textbook section.

The first two sections came from a chapter about reduction-oxidation reactions with the first section introducing the topic and the second section describing reduction potentials. The third textbook describes the structure and function of of adenosine triphosphate (ATP). The fourth textbook section describes the process of glycolysis, the metabolic pathway that extracts energy from glucose.

System Template

You are an educational expert specializing in developing assessment questions tailored to specific requirements. You should always first review carefully the intended learning objectives and other characteristics for a question, and then compose it so that it best meets those requirements. Especially, pay attention to the cognitive skill that the question is expected to target, keeping in mind Bloom's Taxonomy of learning objectives:
Here are the strategies:
- Remembering: Focuses on learners' ability to memorize, recognize, and recall information, facts, details, and terms. - Understanding: Focuses on learners' ability to comprehend the meaning of concepts and principles and explain it in their own words. Related questions should assess learners' ability to interpret, demonstrate, classify, summarize, infer, compare and/or explain.
- Applying: Focuses on learners' ability to apply knowledge (concepts, approaches, principles, techniques, skills) and use it to solve problems or perform tasks in new situations and contexts.
- Analyzing: Focuses on learners' ability to break down information into its constituent parts and identify patterns, relationships, or connections among them.
- Evaluating: Focuses on learners' ability to make judgments about the value or quality of ideas, solutions, or arguments. They can critique, assess, and defend their positions.
- Creating: Focuses on learners' ability to generate new ideas, concepts, or products. They can synthesize information from different sources to create something new.

Figure 3.4: The System Template provides information to the LLM about the six Bloom's Taxonomy levels.

Objective: Remembering

Learning Goal: ME.16 Apply the concept of the "conservation of mass" to metabolism by describing the different forms mass takes as it enters and leaves the cell (e.g. input: reduced molecules like glucose, lipids, proteins, etc. output: oxidized molecules like CO2, H2O etc.).

Question: Which of the following best describes the process of conservation of mass in cellular metabolism?

Options:

A. The transformation of reduced molecules like glucose into ATP and NADH within the cell.

B. The conversion of inputs like glucose, lipids, and proteins into oxidized molecules such as CO2 and H2O as they enter and leave the cell.

C. The continuous recycling of ATP and ADP within the cell for energy transfer.

D. The synthesis of ATP through substrate-level phosphorylation and oxidative phosphorylation.

Correct Answer: B

Explanation: The conservation of mass in cellular metabolism involves the transformation of input molecules like glucose lipids and proteins into oxidized molecules such as CO2 and H2O as they enter and leave the cell. This reflects the principle that mass is conserved during metabolic processes.

Figure 3.5: An example question generated with textbook section 3 using GPT-4o

Using these four sections a total of 64 questions, 16 from each section were generated equally across each Bloom's taxonomy level. Additionally, 4 of the 16 questions in each section were generated according to an inputted learning goal along with the inputted Bloom's level. This set of 64 questions form the basis for the first round of evaluation. A description of questions generated can be seen in table 3.1 and an example of a question generated by the model can be seen in figure 3.5.

The second round of evaluations was conducted only using question's from the fifth textbook section. In this round 16 questions were generated from the fifth textbook section which concentrated on describing pKa values, a measure of a molecule's acidity. Each of these questions were generated with inputted learning goals and an example of a well-structured question from the same class the textbook is used in. Among the 16 generated questions, there are 4 questions for each of of the four considered Bloom's taxonomy levels and among questions of the same taxonomy level, each is generated according to a unique learning goal so that no two questions share the same taxonomy level and learning goal.

As part of our second experiment, our evaluator panel was made up of three experienced biology instructors from UC Davis. These evaluators gave feedback on the 16 questions from section 5. These 16 questions were then revised using their feedback to produce a refined set of questions at with the same taxonomy levels and learning goals. A final survey was given with the 16 revised questions to the same panel of evaluators and two of the original three who gave feedback in the first round left feedback for the revised questions.

3.3 Human Evaluation

The evaluation of the dataset of LLM-generated MCQs is intended to address research questions **RQ2** and **RQ3** concerning the ability of an LLM to generate MCQs corresponding to Bloom's taxonomy levels and the capacity of the model to create questions of a high quality. To measure the quality of questions generated, this study involved human evaluation with two sets of science educators. The first experiment worked with secondary school science teachers in Scotland to evaluate the quality and Bloom's level of questions from the first four textbook sections. The second experiment used the help of three biology instructors from UC Davis to evaluate the quality of questions before and after automatically integrating their feedback.

Survey Items

- Is the question relevant to the respective excerpt?
- Does the question contain correct information?
- Is the question grammatical and well formed?
- Does the question only contain one correct answer?
- Do the wrong answers target misconceptions that students may have?
- Would you use this question in your course?

Figure 3.6: A sample of the questions included in the first questionnaire.

3.3.1 Experiment 1: Evaluation with Secondary School Teachers

As part of the first experiment, a total of 8 secondary school science teachers were asked to rate the quality and Bloom's level of LLM-generated multiple choice questions. Each teacher received 16 MCQs and the textbook section the questions were based on. The teachers were then asked to evaluate each generated question across 10 survey items concerning the MCQ's quality and suitability for the classroom. They were also asked to rate the MCQ's Bloom's taxonomy level from among the 4 levels used to generate MCQs (Remembering, Understanding, Applying and Analyzing).

In addition, 4 of the 16 MCQs that each teacher received were generated using a learning goal. For these MCQs, teachers were also asked to rate whether answering the MCQ helped in progressing towards the learning objective. Figure 3.6 is a list of a few of the evaluation items used to gauge a MCQ's quality. The full list of survey items for each MCQ can be found in appendix B.1.

To create a quality score for each MCQ, responses for survey items concerning the MCQ's quality were coded with values 0 to 2, where 2 represents the highest quality scoring for a question and 0 represents the lowest, with the exception of the question "*Assuming you were teaching the same material, would you use this question in your course?*" for which there were only two possible responses, yes and no. Therefore MCQ's can only receive a score of 0 or 1 for this particular evaluation item. A sample of how responses were coded is shown in 3.7.

To generate an overall quality score for each MCQ, the sum of their response scores for each survey item regarding quality of the MCQ was taken. The highest possible

Coding Questionnaire Response

- Is the question relevant to the respective excerpt?
  0. No, the question is not relevant
  1. The question is relevant but requires knowledge beyond what is included in the text section
  2. Yes, the question is relevant to the excerpt
- Does the question contain enough information to arrive at an answer?
  0. No, it doesn't provide enough information
  1. Somewhat, but additional information would be useful for clarity
  2. Yes, it provides enough information
- Assuming you were teaching the same material, would you use this question in your course?
  0. No
  1. Yes

Figure 3.7: A sample of how questions were coded for quality.

Additional Survey Items in Experiment 2

- The question was generated to match this Bloom's level: [Bloom's Level given to LLM]. If your selected Bloom's level does not match, what could make the question of the desired level?
- What can be changed to make the question be of a higher quality? Please provide directed feedback on what changes could be made for the question to be better.
- When would using this question be most appropriate?

  A. Exam Question

  B. Home Practice Question

  C. Reading Comprehension

  D. Other [Please explain]

- Please explain why you selected the question context you did in the last question.
- Do you have any additional feedback about this question?

Figure 3.8: The additional open response survey items included in the second experiment.

value for this score was 19, which meant that a MCQ received the highest ratings for every question asked about its quality.

3.3.2 Experiment 2: Feedback Integration

The second experiment occurred in two stages. First three university-level biology instructors from UC Davis provided quality evaluations and feedback for question improvement on MCQs. This feedback was then used to automatically revise the MCQs. In the second stage, these same instructors were asked to evaluate the quality of the revised MCQs.

For this experiment, 16 questions were generated from the fifth textbook section which is concentrated on the topic of pKa values. Each of these 16 questions were generated using a specified Bloom's level and a learning goal that was provided within

the textbook [12]. In addition to the survey items that were included in the first experiment conducted secondary school teachers, involving evaluation of the quality and Bloom's level of each MCQ, the second experiment also asked instructors to provide open response feedback to survey items that asked instructors about how questions could be refined. Instructors were also asked to describe in what context (at home, on an exam, etc.) the MCQ would be most appropriate. The additional survey items asked as part of the second experiment are described in the figure 3.8. Slight changes were made between the survey items in experiment 1 and experiment 2. For example experiment 1 included one survey item asking whether distractors were plausible and another survey item asking whether distractors addressed common misconceptions students may have. These two survey items were combined into one in experiment 2. The full list of survey items in experiment 2 can be found in appendix B.

Once feedback was received from our human evaluators, feedback for each MCQ was selectively concatenated from each of the evaluators and used to revise each of the MCQs. Feedback received in the first stage that was overly vague, not relevant or that rewrote the MCQ in its entirety was not considered when revising the MCQs.

In the second stage of this experiment, the same set of instructors received another survey with the 16 newly revised questions. The second survey included identical evaluation and Bloom's level survey items as the first, with an added survey item that included the original form of the question and the feedback that was provided for revision. This survey item asks evaluators '*Does the revised form of this question address the feedback provided?*'. Evaluators were also asked to provide an explanation for their response to this item.

A similar procedure to that described in experiment 1 involving coding the responses of the survey items was used to generate an overall quality score for each of the 16 MCQs in experiment 2 to understand the quality of each MCQ before and after revision.

3.4 Bloom's Level Evaluation with Supervised Models

A series of five traditional machine learning models (Gaussian naive Bayes, support vector machine, logistic regression, random forest and XGBoost) and one BERT-based deep learning model were trained to predict Bloom's taxonomy level classifications following the work set out in [22]. These models were then used to predict the Bloom's level of each of the questions within our LLM-generated dataset. The predictions of these models are compared with the Bloom's levels that our questions were generated

with respect to, to better understand whether the questions generated by the LLM match the Bloom's level that they were instructed on.

3.4.1 Training Datasets

A combination of three datasets were used to train the supervised model. While there are large datasets which map learning objectives to Bloom's taxonomy levels [22], as this research was focused on understanding the Bloom's level of the question stem generated by GPT-4o, the training set was only comprised of question stems. The dataset created in [15] contains a total of 3,397 samples of multiple choice questions. 903 of the questions included a label for their cognitive complexity level according to Bloom's taxonomy. Only these 903 were used within our dataset. The Yahya [43] dataset is a sample of 600 open response questions labelled according to their Bloom's levels. Questions labeled with synthesis and evaluation levels were not used for the combined dataset as questions were only generated by GPT-4o according to the first four levels of Bloom's taxonomy. Additionally, a dataset of 1772 open response questions with their labeled Bloom's levels was collected from Kaggle [36]. Once again, only questions labelled with the first four Bloom's taxonomy levels was retained for our combined dataset, resulting in a total of 1171 questions used. These datasets were collected and used for training the machine learning models as there is not currently an available dataset of LLM-generated MCQs pre-labelled with their Bloom's levels to train on.

Questions were all converted to lowercase and whitespace was removed. Duplicate questions were removed from the combined dataset resulting in a total of 2436 questions that were used for training supervised models.

Bloom's Level	**Hadifar**	**Yahya**	**Kaggle**	**Combined Dataset**
Remember	660	100	300	1024
Understand	114	100	300	509
Apply	110	100	300	507
Analyze	19	100	271	396
Total Questions	**903**	**400**	**1171**	**2436**

Table 3.2: Datasets from [15], [43], and [36] were combined into one dataset used for training supervised models.

3.4.1.1 Data Preprocessing

To pre-process data for the machine learning models, the approach of [22] was followed, whereby 1,000 unigram, 1,000 bigram and 1,000 TF-IDF features were computed. In addition, 119 features were generated from the LIWC 2022 dictionary, a dictionary that reflects frequency of different psychologically meaningful word groups. [6] Therefore, our input data as a combination of the n-gram features and the LIWC features has a total of 3,119 features.

For the BERT-based model, input question data was first tokenized using the BERT tokenizer to convert the questions into sequences of tokens compativle with the BERT model's vocabulary. The tokenized sequences were then converted into PyTorch tensors that were combined into a TensorDataset.

3.4.2 Training and Prediction

To apply supervised learning models to classifying the LLM-generated questions, five machine learning models were used: Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF) and XGBoost (XGB). Hyperparameter optimization using grid search with cross-validation was used for each model where the goal was to identify the best hyperparameters that maximize the F1-macro score. A full set of the hyperparameters used for optimization can be found in table 3.3.

A pre-trained BERT (Bidirectional Encoder Representations from Transformers) model was modified to classify each question according to Bloom's levels [9]. In this research, the model used is a fine-tuned version of 'bert-base-uncased', which contains approximately 110 million parameters. The core architecture of the BERT model is composed of several key components: the embeddings layer, which converts the tokenized text into embeddings, and the transformer encoder, which includes 12 layers containing a self-attention mechanism and a feedforward neural network. Additionally, the pooler and classification head pass the pooled output through the classification head to generate predictions regarding the Bloom's level. The model was fine-tuned using the 'BertForSequenceClassification' class which includes the pre-trained BERT model and an extra classification head for sequence classification tasks. The model was trained for 5 epochs with a batch size of 32 and a learning rate of 2e-5 and was optimized using the AdamW optimizer.

For both the BERT model and the traditional machine learning models, the dataset was split into training and validation sets in an 80:20 ratio. For the BERT model, the

Model	**Hyperparameters**
Naive Bayes	`var_smoothing` = {1e-8, 1e-9, 1e-10}
SVM	`C` = {0.1, 1, 10, 100} `gamma` = {'scale', 'auto'} `kernel` = {'linear', 'poly', 'rbf'}
Logistic Regression	`penalty` = {'l1', 'l2', 'none'} `C` = {0.1, 1, 10} `solver` = {'saga', 'liblinear'} `tol` = {0.01, 0.001, 0.0001} `max_iter` = {200, 500}
Random Forest	`n_estimators` = {50, 100, 250} `max_depth` = {None, 5, 10} `max_features` = {'auto', 'sqrt'} `min_samples_split` = {2, 5, 10} `min_samples_leaf` = {1, 2, 4} `bootstrap` = {True, False}
XGBoost	`gamma` = {0.1, 0.5} `learning_rate` = {0.1, 0.5} `max_depth` = {5, 7, 10} `n_estimators` = {50, 100}

Table 3.3: Hyperparameters tested for different machine learning models.

training set was used to fine-tune the model while the validation set was used to monitor the model's performance and prevent overfitting. For the machine learning models, the validation set was used to test the accuracy and F1 score of each model.

In order to generate predictions on the LLM-generated questions, features for the set of generated questions were created according to the same procedures as input features. Once models were trained on the original training set of data, predictions for the Bloom's levels of the LLM questions were generated.

Chapter 4

Results and Discussion

In this chapter, the results of the two experiments involving human evaluators and the results of Bloom's alignment involving supervised models will be described. The implications of these results regarding the capacity of LLMs to generate MCQ automatically will also be discussed.

4.1 Human Evaluation

Human evaluation of the LLM-generated MCQs was conducted in experiment 1 and experiment 2. These two experiments quantified the quality of MCQs generated and used human evaluators to predict the alignment of MCQs with Bloom's levels. This section will discuss the results of first the quality assessment of the MCQs, then the alignment of the MCQs with assigned Bloom's levels.

4.1.1 Quality Evaluation

4.1.1.1 Experiment 1

The results of the overall quality scores in experiment 1 distributed across the four considered Bloom's levels can be seen in table 4.1. As seen in the figure 4.1, the MCQs generated according to the Application Bloom's level were judged by human raters as having an overall higher average and median quality rater than questions generated at all other Bloom's levels. However, the difference between the average quality rating for each Bloom's level was not significantly different.

The Bloom's level at which GPT-4o was instructed to generate a question does not seem to have significantly impacted the overall quality of each question. Teachers were

asked to judge the Bloom's level of each MCQ without seeing the Bloom's level the LLM used to generate the question. When observing the average quality rating for each Bloom's level as evaluated by the teachers, the average quality rating shows a much larger separation between evaluated levels. MCQs that were believed by teachers to be at the Apply level of Bloom's taxonomy received an average quality score of 16.48 while questions that were perceived to be at the lowest Bloom's level of Remember received an overall score of 14.58. These results can be seen in table 4.2. Therefore, while the Bloom's levels at which a question was generated on does not seem to affect very much the overall quality of the question, teachers do tend to rate questions that they believe are of a higher Bloom's level with higher quality scores than questions they believe are from a lower Bloom's level.

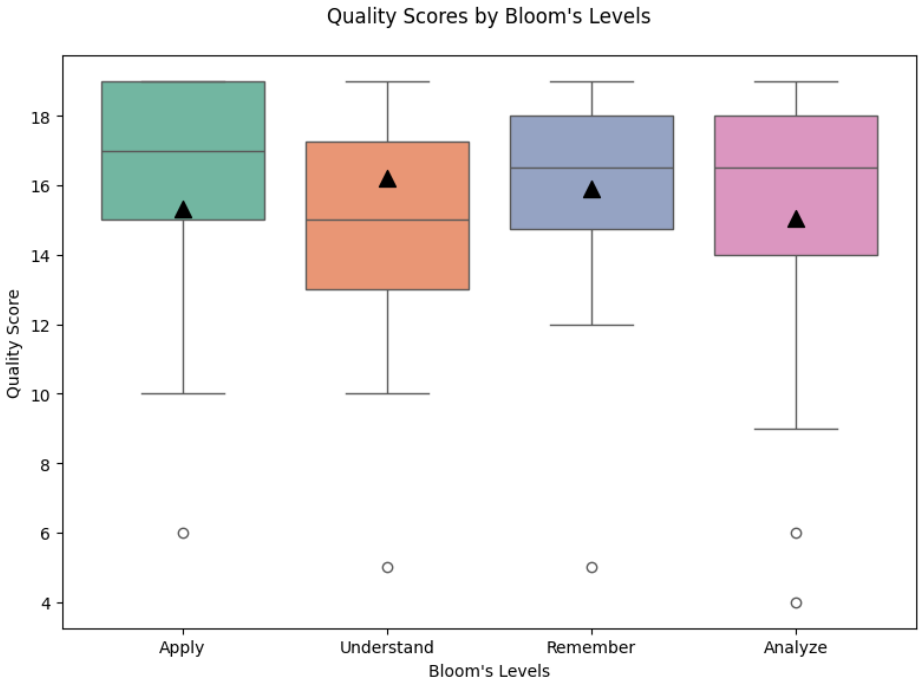

Figure 4.1: The distribution of question quality across Bloom's level in the first survey.

Bloom's Level	**Mean**	**SD**	**Median**
Analyze	15.31	3.80	16.50
Apply	16.22	2.93	17.00
Remember	15.91	2.89	16.50
Understand	15.03	3.13	15.00

Table 4.1: Quality score distributions on Bloom's levels in the first survey.

On average, the MCQs in experiment 1 obtained an average quality score of 15.62. MCQs were most likely to be rated lower on the survey item which asks teachers to rate whether the distractors in the MCQ address common misconceptions that students may have 4.3. On all other measures of quality where 2 points were available, the MCQs on average scored between 1 or 2 which meant that for most measures of quality, the MCQs are being ranked on average between the medium quality responses and the high quality responses.

Predicted Bloom's Level	**Mean**	**SD**	**Median**
Analyze	18.00	NaN	18.00
Apply	16.48	2.44	17.00
Remember	14.58	3.90	16.00
Understand	16.22	2.35	17.00

Table 4.2: The distribution of quality scores across Bloom's levels as predicted by secondary teachers.

Quality Measure	**Average**	**SD**
Relevant	1.727	0.557
Correct Information	1.781	0.546
Grammatical	1.656	0.608
Enough Information	1.805	0.518
Correct Answer Marked	1.891	0.439
One Correct Answer	1.961	0.231
Distinct Options	1.797	0.476
Plausible Distractors	1.383	0.915
Misconceptions Addressed	0.969	0.813
Usable	0.648	0.479

Table 4.3: Averages and Standard Deviations of measures of MCQ quality in experiment 1.

4.1.1.2 Experiment 2

In experiment 2, we evaluated the capacity of the LLM to incorporate instructor feedback to generate revised versions of questions. Three biology instructors from UC Davis

initially rated 16 questions generated from the pKa section of a textbook. Questions in this second experiment were all generated with the additional instruction to write a MCQ that assess a particular learning goal such as *Define and correctly use the acid dissociation constant, Ka.*

In addition to providing ratings, the instructors offered specific feedback on how to enhance the quality of the questions and better align them with the intended Bloom's taxonomy level. This feedback was subsequently used, along with the original questions, to prompt GPT-4o to revise the questions accordingly. Examples of the feedback included comments such as, "*The selected correct answer choice is wrong—though the explanation is correct and does a nice job*," "*The punctuation in the answers made this question difficult to figure out (I found adding semi-colons helped)*," and "*To make the question an 'Analyzing' question, I would require students to select the correct dissociation behavior and explanation*." We then compared the quality scores of the original set of 16 questions to those of the revised questions to assess the impact of the revisions on question quality. As in experiment 1, the responses to survey items regarding a MCQ's quality were coded on a scale from 0 to 2 with 0 being a response that indicates low quality for the measure and 2 indicating a MCQ with high quality for the metric.

Quality Measure	**Initial Average**	**Revised Average**
Relevant	1.938	2.000 (↑ 3.23%)
Correct Information	1.833	2.000 (↑ 9.09%)
Grammatical	1.917	1.844 (↓ -3.80%)
Enough Information	1.702	1.906 (↑ 11.99%)
Correct Answer Marked	1.792	2.000 (↑ 11.63%)
One Correct Answer	1.938	1.938 (- 0.00%)
Distinct Options	1.813	1.875 (↑ 3.45%)
Plausible Distractors	1.625	1.813 (↑ 11.54%)
Learning Goal Progress	1.813	1.813 (- 0.00%)

Table 4.4: Comparison of question quality measures before and after revisions.

As seen in table 4.4, on almost all measures of quality, the MCQS scored very high, with the average for each measure of quality scores greater than 1.5, indicating that MCQs tended to be rated with the highest level of quality for each metric. In addition, for all but one metric, the MCQs that were revised with feedback achieved

higher ratings than the originally generated questions. In the original set of questions, 77.1% of responses to the question *"Is the MCQ free from obviously-wrong options? Are all distractors plausible answers that target student misconceptions?"* received the highest rating of *"Yes there are no obviously wrong options and all distractors target common misconceptions"* and in the set of revised questions, 90.6% of responses gave the highest rating. The percentage of responses that rated the question as *"relevant to the respective excerpt"* increased from 93.8% in the original question set to 100% in the revised question set. For the question, *"Is there a correct answer listed in the options and is the option marked "correct" actually correct?"*, 83.3% of responses in the first round replied "*Yes, there is a correct answer and it is marked as the correct answer*". This rating increased to 100% after the questions were revised. Additionally, while 78.7% of responses replied to "*Does the MCQ provide enough information to arrive at an answer?* with *Yes, it provides enough information* for the first set of MCQs, this percentage increased to 93.8% for questions revised with feedback. These results are displayed in figure 4.2.

While most questions about the quality of the MCQs saw a significant increase in their ratings after the questions were revised with feedback, there are a few notable exceptions. The original set of MCQs before revision received an average score of 1.8 on the question of if the MCQ "*helps in progressing towards*" its given learning goal. This score did not change after revision. In addition, there was no change in quality after revision on ratings of if "*the question only [contains] one correct answer*". The average rating of whether the MCQs were "*grammatical and well formed"* went down from 1.9 to 1.8 after revision.

In addition to evaluating the quality of the MCQs, it is essential to determine whether instructors would be inclined to use these questions in their courses, either as reading comprehension tools or exam questions. As seen in figure 4.3 initially 70.8% of responses by instructors indicated they would use the original questions in their course. Following revisions informed by instructor feedback, this percentage increased to 84.4%, suggesting that the revisions made the questions more suitable for instructional use.

In the second round of surveys with revised questions, instructors were also provided the original set of questions along with the feedback used to revise each question. They were then asked to rate whether the revised form of the question addressed the feedback provided. As seen in figure 4.4, 56.2% of responses to this question indicated that the revision significantly addressed the feedback with 25.0% of responses stating that the

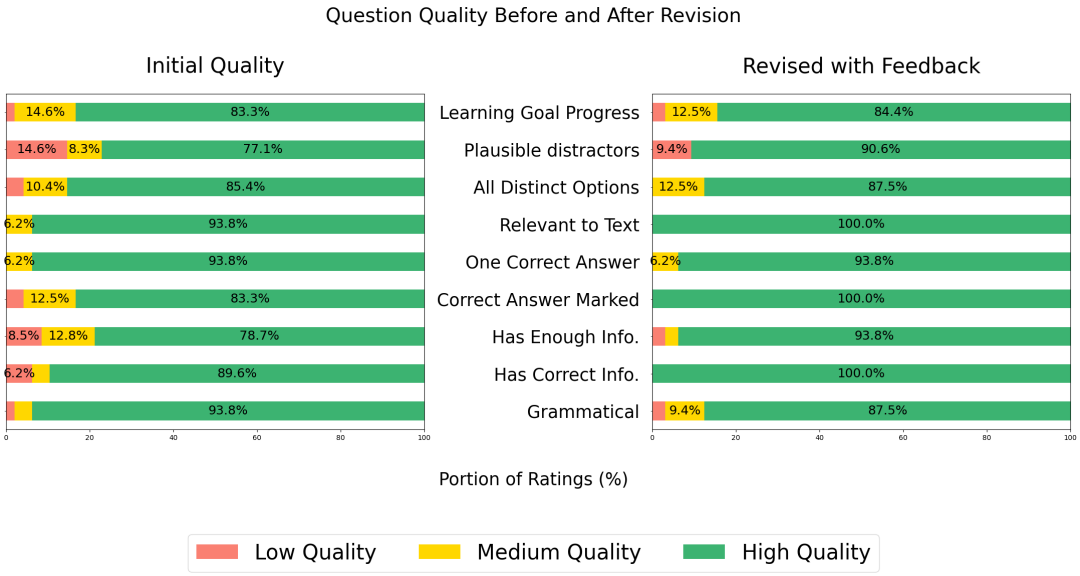

Figure 4.2: The change in quality evaluation after feedback.

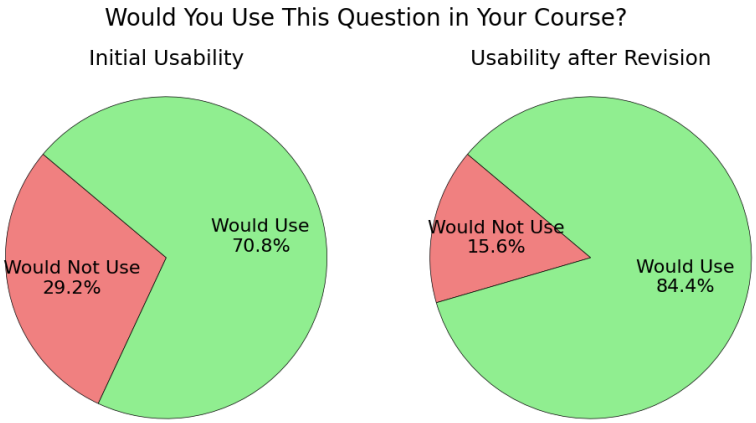

Figure 4.3: Questions were more likely to be rated as suitable for instructional use after being revised with feedback

revision partially addressed the feedback and 18.8% of responses stated that the revised form did not address the provided feedback.

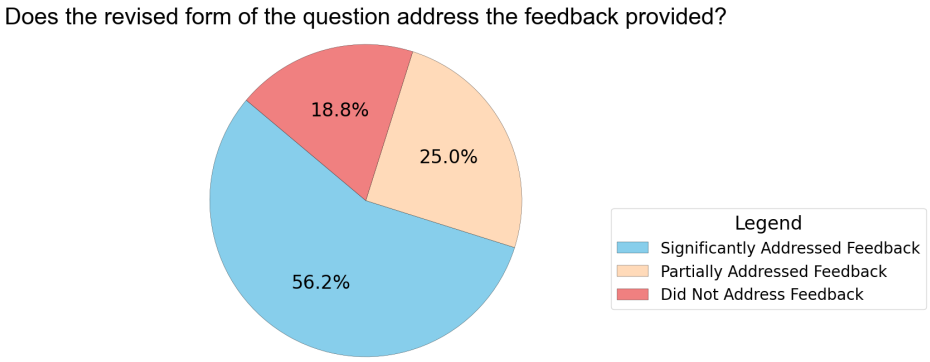

Figure 4.4: A majority of responses indicated that the revised question ”significantly addressed” the feedback provided.

4.1.2 Bloom’s Level Alignment

In addition to quality, MCQs were investigated on whether the Bloom’s levels they were generated to address by the LLM match the Bloom’s levels that teachers judged the MCQs to be.

In experiment 1, teachers predicted the Bloom’s levels of MCQs with an accuracy of 38.28% with respect to the Bloom’s levels that the LLM was instructed to follow. The Kappa κ (interrater reliability) score between the GPT-generated levels and the teacher’s predicted Bloom’s levels is .177 which indicates none to slight agreement [25]. Figure 4.5 presents a Sankey diagram depicting at which Bloom’s levels teachers judged MCQs to be, based on the Bloom’s level at which the LLM generated the MCQ.

While there is much disagreement between the Bloom’s labels the LLM gave to each MCQ and the Bloom’s levels teachers indicate the MCQs to be, there is also much disagreement among teachers. In experiment 1, for MCQs that were rated by more than one teacher, we computed the interrator reliability score between teachers using Fleiss’ Kappa [19]. This interrator reliability score was computed to be -0.027 indicating a very poor level of agreement among teachers on the Bloom’s level of each MCQ.

In experiment 2, involving improving questions with feedback, accuracy and κ scores were calculated between the LLM-generated levels and the Bloom’s levels as judged by instructors for the 16 questions involved in this experiment. In the first survey, before revision with feedback, instructor predicted Bloom’s levels had an accuracy of

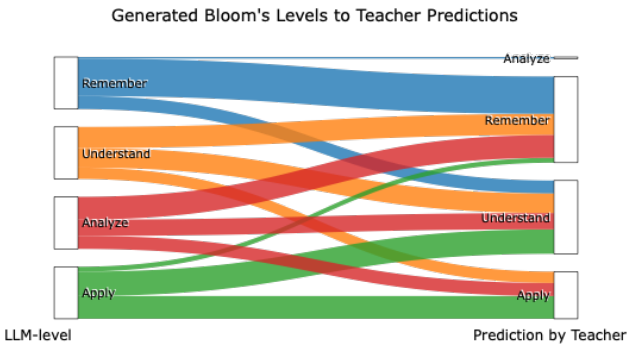

Figure 4.5: The correspondence of Bloom's levels as generated by the LLM to the levels predicted by teachers.

56.35% and a κ score of 0.417 indicating moderate agreement between the generated Bloom's levels and the predicted levels of these questions [25]. After revision, the accuracy of Bloom's levels increased to 62.5% and the κ score increased to 0.5, still indicating moderate agreement. This provides some evidence that not only did revision of the MCQs with feedback increase the quality of the questions, but it also seems to have made the questions better aligned with the Bloom's levels the LLM was instructed to generate. As seen in figure 4.6, after revision of MCQs, instructors judged MCQs that were generated at the Remember and Analyze levels more accurately according to the Bloom's level used to generate them. While instructors had 25% accuracy on judging MCQs at the Analyze Bloom's level before revision, after the revision of MCQs, the accuracy for MCQs generated at the Analyze level increased to 50%. Questions generated at the Remember level were also more likely to be judged by instructors accurately as the correspondence between the LLM-generated labels and the instructor judged labels increased from 58.3% to 87.5%.However, instructors were less likely to judge questions generated at the Apply and Understand levels accurately after revision as these values decreased from 100% to 87.5% and from 41.7% to 25% respectively. Interestingly, after feedback revision, questions generated at the Apply level were slightly more likely to be identified by instructors as matching the Analyze Bloom's level which may indicate that these questions increased in cognitive complexity.

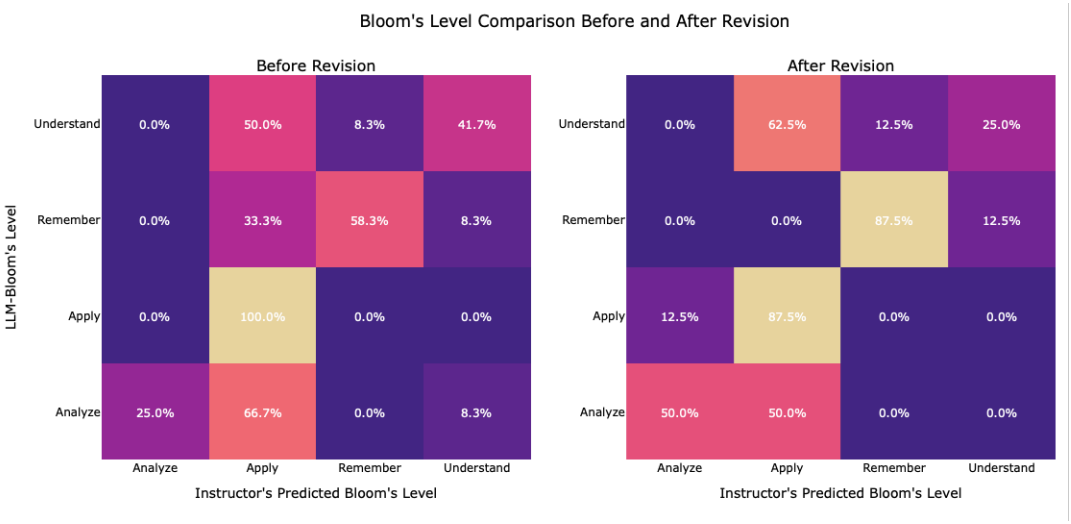

Figure 4.6: The correspondence of labels the MCQs were generated on against the Bloom's level prediction provided by instructors before and after revision.

4.2 Supervised Evaluation

As stated previously we used 6 models to generate predictions for Bloom's taxonomy levels on a set of 80 LLM-generated MCQs (64 from the survey involving secondary school teachers, 16 from the survey with UC Davis instructors) by the process outlined in 3.4. Once trained using set of questions that are labeled according to their Bloom's levels, these models were used to predict the Bloom's levels of the LLM-generated questions.

As presented in table 4.5, the BERT model achieved an accuracy on its testing set of 0.8709. The logistic regression, random forest and XGBoost models all achieved accuracies above 0.80 on their testing sets while the naive bayes model had an accuracy of 0.59 on its testing set and the the SVM model had an accuracy of 0.75 on its testing set. The fact that these machine learning models perform with this level of accuracy on their testing sets, which come from the original dataset collected, indicates that these models are capable of predicting Bloom's levels when on questions that are similar to their training dataset.

We then moved to test the prediction capabilities of the trained models on the MCQs generated by the LLM. The results of this are shown in table 4.6. The model with the highest level of agreement, the refined BERT model, reported an accuracy of 0.38 and a κ score of 0.18. While the BERT model performs much better at predicting the levels of the LLM-generated questions compared to the five machine learning models, the difference between BERT's high accuracy on its testing set (0.87) and its much lower

Model	Accuracy on Test Set
BERT	**0.87**
Naive Bayes	0.59
Support Vector Machine	0.75
Logistic Regression	0.80
Random Forest	0.81
XGBoost	0.83

Table 4.5: Accuracy of each of the models on their testing set following training.

accuracy on the LLM-generated questions is significant. Other than the BERT model, only XGBoost achieved over an accuracy of 0.30 on predicting the Bloom's levels of the MCQs and κ values for all models indicate slight to no agreement between the labels given to the MCQs by the LLM and those predicted by the supervised models.

Model	Accuracy	Kappa Score	F1 Score
BERT	**0.38**	**0.18**	**0.34**
Naive Bayes	0.25	0.03	0.15
Support Vector Machine	0.29	0.05	0.21
Logistic Regression	0.28	0.03	0.21
Random Forest	0.26	0.02	0.13
XGBoost	0.31	0.08	0.27

Table 4.6: Performance metrics of different supervised models on the generated MCQ set.

4.3 Discussion

In experiment, the LLM-generated questions received relatively high-quality ratings, with each Bloom's level achieving an overall score of 15.03 or higher out of a possible 19 points. The Bloom's level at which the LLM generated the question had only a minor impact on the average quality score, with a difference of just 1.19 points between the lowest scoring level, Understand, and the highest scoring level, Apply. The fact that the different levels at which the LLM was instructed to generate the MCQs does not significantly impact the overall quality score of a MCQ suggests that GPT-4o is able to generate high-quality questions at a range of cognitive complexities. While past

research found that the quality of questions generated by LLMs degraded with questions generated at higher Bloom's levels [5], the quality of our MCQs do not seem to change significantly when high Bloom's levels are considered. By injecting knowledge in the form of a description of Bloom's taxonomy and instructions about how to create a well formed MCQ, the LLM was able to perform similarly well, even when given different Bloom's levels.

Overall, the results of revising MCQs with teacher feedback proved to be successful. Almost all questions involving the question's quality saw an increase in their average rating after integrating feedback. In addition, 81.2% of responses indicated that the revised question had at least partially addressed the feedback provided. Questions also became more suitable for classroom usage after revision with more instructors stating that they would use the MCQ in their course.

These results indicate a promising path for improving the quality of LLM-generations using human provided feedback from experts. A model of artificial intelligence that requires human interaction is considered "human-in-the-loop" [28]. Our method of revision of feedback creates a step of additional human interaction with the LLM, integrating human-in-the-loop principles, as the human must reflect on the question created by the LLM and provide meaningful feedback on how to increase the quality of the question. While instructors themselves did not as a part of this project, use the EduGenie application themselves to generate revised questions, as this work was done by the researcher, the next step to more quickly integrate instructor feedback would be to make this generation and revision process an easy experience so that educators could use the application themselves to directly and immediately provide feedback.

This research did find difficulties in aligning the Bloom's levels of generated MCQ correctly. In the first survey with secondary school teachers, the low accuracy of predicted Bloom's levels (38.28%) indicates that the levels at which the questions were intended to be generated were not easy to identify. Predicted Bloom's levels were more accurate in experiment 2 starting from 56.35% and increasing after revision with feedback to 62.5%. These low levels of accuracy between the Bloom's level that the LLM was intended to generate and the predicted level by instructors seems to suggest that the LLM is not very accurate in producing questions at a specified Bloom's taxonomy level. However, the high level of disagreement on the predictions Bloom's levels of each MCQ among teachers, with a Fleiss' Kappa value of -0.027, may put into question treating the prediction of Bloom's levels by teachers as ground truth. This low level of agreement suggests that among themselves teachers do not always

agree to which Bloom's level a question belongs. However, when teachers did disagree, they often did so by each identifying neighboring Bloom's levels (remembering and understanding, understanding and applying or applying and analyzing), and rarely identified the same question as differing by more than one Bloom's level. Out of 120 paired ratings of Bloom's levels where 2 teachers rated the same question, only 8 ratings between 2 teachers differed by more than one Bloom's level. Thus for 93.3% of ratings teachers either agreed on the Bloom's level of a question or differed in the level they identified by only one level.

When applying supervised models to generate predictions on our set of MCQ, although the supervised models performed well on their training sets, their accuracy predicting the the Bloom's levels of our generated questions was significantly lower. The difference between the models performance on the training sets and our MCQs has at least two potential explanations. The first is that the LLM has done a poor job generating MCQs at the correct Bloom's levels. The second explanation is that the set of questions which the models were trained on differ from the LLM-generated MCQs so that the models are not accurately able to predict the labels of these questions. There are some significant differences between the training set of questions and our MCQs that may point to this second explanation. The mean number of words per question in the training dataset is 66, while the mean number of words in the LLM-generated question set is more than twice as many at 133.5. In addition while the Hadifar [15] dataset that was used contains only multiple choice questions, the dataset from Kaggle [36] and Yahya [43] both contain open response questions. Each of these three dataset that were combined also cover a large number of different class subjects while our LLM-generated set of questions only includes questions about biology. To develop a supervised model that can perform better on such a set of MCQs, it would be best to use larger labeled datasets only containing MCQs in the subject of interest.

As a part of this research project, feedback was also received from instructors about the different contexts in which these generated questions may be useful. Questions that tested basic factual knowledge, or that were at a lower Bloom's level were said to be a better fit for home practice questions while higher Bloom's level questions were better suited to exams.

One piece of common feedback on the generated-MCQs were that they were often *"too long and wordy"* and needed to be *"more concise"*. This is not just an inconvenience in that reading the questions can take longer, but questions that are longer and more difficult language can put students *"at a disadvantage based on their reading/English*

skills rather than their biology/chemistry skills" as one instructor noted. In order to address the wordiness of generated questions, going forward it will be important to include more requirements on concise language within the initial prompt and within any provided feedback.

When asked about their opinions on using AI to generate educational content, teachers on the whole seemed to be responsive to the topic with some caveats. Many teachers pointed to the current need to spend significant time checking the work done through automatic generation, with one teacher stating *"I still find I am doing a lot of checking after"* since as another suggested *"Any work generated has to be checked thoroughly for mistakes and to make sure it makes sense"*. It is also essential that we integrate mechanisms within the generation process for personalisation as any questions or educational content generated must *"[fit] to the specific curriculum requirements"* set out by each class. While AI generated questions still require significant time to check and revise before using in a classroom setting, teachers within our survey were generally positive to using AI with one stating *"There are many nuisances to generating questions but AI gives a great basis to start from"* and another gave praise saying *" I do feel that AI can be a really useful tool and would be a great asset in generating questions/ activities to support pupils. With increased teacher workload, I feel that this would be a welcomed addition to aid teachers in their development of learning resources"*. One teacher who is more hesitant to using AI to generate content also seemed to be compelled by the research project writing *"I didn't think they [AI-generated questions] would be useful but having seen these questions it might be useful"*.

Overall this research shows that the multiple choice questions generated in this study were of a high enough quality on average to be considered for usage within the classroom. Revising questions using written feedback from class instructors was successful in improving the quality of questions generated. There are still significant problems however, posed by aligning the Bloom's levels of questions that the LLM is instructed to generate with the Bloom's levels predicted by teachers and by supervised models.

Chapter 5

Conclusions

5.1 Revisiting Research Questions

The three research questions that this research project sought to address are as follows:

1. How can the process of creating multiple choice questions be automated to support lecturers in teaching?
2. Can LLMs generate multiple choice questions that correspond to specified Bloom's Taxonomy levels?
3. Are multiple choice questions generated by large language models of a high enough quality to be usable within a classroom setting?

To automate generating multiple choice questions for exams, this project built on the work of [5], adapting an application built in Python that manages prompting and parsing responses from the GPT-4o model. The multiple choice questions generated required a Bloom's taxonomy level to be specified and a textbook section based on which questions should be generated to be provided.

The Bloom's taxonomy levels of the created questions were tested by human prediction from a set of secondary school science teachers and biology instructors from UC Davis. The Bloom's taxonomy levels of our generated questions were also predicted by a series of supervised machine learning and deep learning models. In both the supervised learning approach and the human evaluation approach, it was shown that the LLM does not currently generate questions that correspond with high accuracy with the Bloom's levels predicted by either our models or human evaluators. In particular, our research found that questions generated according to the Bloom's level Analyze were the least likely to be judged at this level by human evaluators.

The questions generated by GPT-4o did receive high quality ratings from human evaluators based on a survey that asked to evaluate the quality of the generated questions on a number of metrics. Revision of questions using feedback provided by instructors also proved to be a powerful tool in enhancing the quality of the MCQs and teachers were more likely to state that they would use the questions in their course after questions were revised using provided feedback.

5.2 Limitations

One major limitation of this project is that only one LLM, GPT-4o, was considered. This decision was due to time constraints and due to financial costs.

Due to the short period of the project and the number of human evaluators who were available to provide feedback, it would not have been possible to significantly increase the size of our dataset of generated questions and receive evaluations for every one.

Additionally new models are a more expensive investment. At time of writing, GPT-4o costs $5.00 per 1 million input tokens and $15.00 per 1 million output tokens while the more powerful model, GPT-4, costs $30.00 per 1 million input tokens and $60.00 per 1 million output tokens, which is 6 times more expensive than the more basic model [32].

Another limitation of the project is that only text based questions were considered. In biology, it is common that questions include a visualisation or diagram where the question asks a student to interpret something from the diagram. To more closely mirror questions, it would be worthwhile to explore multi-modal models that could generate questions involving both texts and images.

The supervised model approach was also limited in its capacity due to the datasets of questions with Bloom's labels available. The ideal dataset would have only multiple choice questions focused on biology topics so that the training data would match the LLM-generated questions that was tested on. However, due to the availability of datasets, the training data was a mix of multiple choice and open response question stems from a wide variety of subjects.

5.3 Future Work

The natural next extension of this work would be to make the application used for generating questions available and accessible to teachers and educators. This would

also have to include creating a quality user experience for educators who would like to use the application because the current way to revise questions using feedback is not very intuitive.

This project used human provided feedback to refine and enhance question quality. While human input is important, it may be possible to also generate and incorporate feedback automatically in the process by using a process of self-refinement whereby an LLM generates an initial output, then the same LLM provides feedback on the output and uses the feedback to refine itself [24]. This would allow for the outputted questions to be revised automatically at test time before receiving feedback from humans. There is an additional opportunity to improve the feedback process by converting feedback provided by instructors to principles that can continuously be used when generating new MCQs [34]. For example when a teacher provides feedback that a question should be concise and not overly wordy, the model could convert this into a principle to only create concise questions which can be used every time a question is generated. As it stands, many teachers see the downside to needing to spend significant time revising the questions generated by LLMs and future work should focus on how to further automate the feedback process.

To enhance the process of question generation, multimodal models could be explored that would allow for multiple choice questions involving images to be created. Currently research on multimodal models have shown an ability to generate questions based on a provided image [33], suggesting a potential for a version of MCQ-generation where a teacher could provide the model a diagram that can then be used to generate a question about the diagram.

5.4 Conclusion

This paper explores the capacity of GPT-4o to generate MCQs aligned on specified Bloom's taxonomy levels and assessed their quality for potential use in classroom settings. The findings indicate that while the LLM is capable of generating high-quality MCQs there are challenges in aligning these questions with the intended Bloom's levels. Despite these challenges, the overall quality of the generated MCQs was rated highly by teachers.

While GPT-4o has demonstrated its capacity to generate high quality multiple choice questions there remains significant areas for improvement. Future work should focus on enhancing the feedback process, exploring multimodal question generation and

making the tool accessible and user-friendly for educators. Other models should also be explored and tested for their ability to generate educational content using similar metrics. With these advancements, LLM-generated questions can be a valuable resource in supporting teachers and improving learning experiences for students.

Bibliography

[1] Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura, 2016.

[2] Katie Bainbridge, Candace Walkington, Armon Ibrahim, Iris Zhong, Debshila Basu Mallick, Julianna Washington, and Rich Baraniuk. A case study using large language models to generate metadata for math questions, 2023.

[3] Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. Multiple-choice question generation using large language models: Methodology and educator insights. In *UMAP 2024 - Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 584–590. Association for Computing Machinery, Inc, 6 2024.

[4] Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. Distractor generation for multiple-choice questions with predictive prompting and large language models. 7 2023.

[5] Ariel Blobstein, Daniel Izmaylov, Tal Yifal, Michal Levy, and Avi Segal. Angel: A new generation tool for learning material based questions and answers. Presented at the NeurIPS 2023 Conference, 2023.

[6] Ryan L. Boyd, Ami Ashokkumar, Sarah Seraj, and James W. Pennebaker. *The development and psychometric properties of LIWC-22*. University of Texas at Austin, Austin, TX, 2022.

[7] Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Øistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. On the application of large language models for language teaching and assessment technology. 2023.

[8] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. Toward personalized xai: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298, 9 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 10 2018.

[10] Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, Christopher Bogart, Eric Keylor, Can Kultur, Jaromir Savelka, and Majd Sakr. A comparative study of ai-generated (gpt-4) and human-crafted mcqs in programming education. In *ACM International Conference Proceeding Series*, pages 114–123. Association for Computing Machinery, 1 2024.

[11] Sabina Elkins, Ekaterina Kochmar, Jackie C. K. Cheung, and Iulian Serban. How useful are educational questions generated by large language models? 4 2023.

[12] Marc Facciotti. Introductory Biology (Facciotti), mar 30 2019. [Online; accessed 2024-08-07].

[13] Mark J. Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6):1082–1116, 2017.

[14] Aytac Gogus. *Bloom's Taxonomy of Learning Objectives*, pages 469–473. Springer US, Boston, MA, 2012.

[15] Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. Eduqg: A multi-format multiple-choice dataset for the educational domain. *IEEE Access*, 11:20885–20896, 2023.

[16] Kevin Hwang, Sai Challagundla, Maryam M Alomair, Lujie Karen Chen, and Fow-Sen Choa. Towards ai-assisted multiple choice question generation and quality evaluation at scale: Aligning with bloom's taxonomy, 2024.

[17] Kazem Jahanbakhsh. Beyond hallucination: Building a reliable question answering explanation system with gpts, 2024.

[18] Archana Praveen Kumar, Ashalatha Nayak, K. Manjula Shenoy, Shashank Goyal, and Chaitanya. A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, 225, 9 2023.

[19] Laerd Statistics. Fleiss' kappa in spss statistics, 2024.

[20] Daniel Leiker, Sara Finnigan, Ashley Ricker Gyllen, and Mutlu Cukurova. Prototyping the use of large language models (llms) for adult learning content creation at scale, 2023.

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. 5 2020.

[22] Yuheng Li, Mladen Rakovic, Boon Xin Poh, Dragan Gaševic, and Guanliang Chen. Automatic classification of learning objectives based on bloom's taxonomy. In *Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022*. International Educational Data Mining Society, 2022.

[23] Chenyang Lyu, Minghao Wu, and Alham Fikri Aji. Beyond probabilities: Unveiling the misalignment in evaluating large language models. 2 2024.

[24] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback.

[25] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

[26] Hunter Mcnichols, Wanyong Feng, Jaewook Lee, Alexander Scarlatos, Digory Smith, Simon Woodhead Eedi, and Andrew Lan. Automated distractor and feedback generation for math multiple-choice questions via in-context learning, 2023.

[27] Niklas Meißner, Sandro Speth, Julian Kieslinger, and Steffen Becker. Evalquiz-llm-based automated generation of self-assessment quizzes in software engineering education. 2024.

[28] Bahar Memarian and Tenzin Doleck. Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (er) analysis. *Computers in Human Behavior: Artificial Humans*, 2(1):100053, 2024.

[29] Abdul Momen, Mansoureh Ebrahimi, and Ahmad Muhyuddin Hassan. Importance and implications of theory of bloom's taxonomy in different fields of education. In Mohammed A. Al-Sharafi, Mostafa Al-Emran, Mohammed Naji Al-Kabi, and Khaled Shaalan, editors, *Proceedings of the 2nd International Conference on Emerging Technologies and Intelligent Systems*, pages 515–525, Cham, 2023. Springer International Publishing.

[30] Benjamin D Nye, Dillon Mee, and Mark G Core. Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns, 2023.

[31] Andrew M Olney. Generating multiple choice questions from a textbook: Llms match human performance on most metrics, 2023.

[32] OpenAI. Openai api pricing, 2024. Accessed: 2024-08-19.

[33] Alkesh Patel, Akanksha Bindal, Hadas Kotek, Christopher Klein, and Jason Williams. Generating natural questions from images for multimodal assistants. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2270–2274. IEEE, 2021.

[34] Savvas Petridis, Ben Wedin, James Wexler, Aaron Donsbach, Mahima Pushkarna, Nitesh Goyal, Carrie J. Cai, and Michael Terry. Constitutionmaker: Interactively critiquing large language models by converting feedback into principles. 10 2023.

[35] Anja Philipp and Mareike Kunter. How do teachers spend their time? a study on teachers' strategies of selection, optimisation, and compensation over their career cycle. *Teaching and Teacher Education*, 35:1–12, 10 2013.

[36] Dinesh Sheelam. Bloom's taxonomy dataset, 2023.

[37] Pragnya Sridhar, Aidan Doyle, Arav Agarwal, Christopher Bogart, Jaromir Savelka, and Majd Sakr. Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives, 2023.

[38] Andrew Tran, Kenneth Angelikas, Egi Rama, Chiku Okechukwu, David H. Smith, and Stephen MacNeil. Generating multiple choice questions for computing courses using large language models. In *Proceedings - Frontiers in Education Conference, FIE*. Institute of Electrical and Electronics Engineers Inc., 2023.

[39] Torrey Trust, Jeromie Whalen, and Chrystalla Mouza. Editorial: Chatgpt: Challenges, opportunities, and implications for teacher education, 2023.

[40] Abdul Waheed, Muskan Goyal, Nimisha Mittal, Deepak Gupta, Ashish Khanna, and Moolchand Sharma. Bloomnet: A robust transformer based model for bloom's learning outcome classification.

[41] Qiao Wang, Ralph Rose, Naho Orita, and Ayaka Sugawara. Automated generation of multiple-choice cloze questions for assessing english vocabulary using gpt-turbo 3.5. 3 2024.

[42] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. 6 2023.

[43] Anwar Yahya. Bloom's taxonomy cognitive levels data set, 01 2011.

[44] Nikki L.Bibler Zaidi, Karri L. Grob, Seetha M. Monrad, Joshua B. Kurtz, Andrew Tai, Asra Z. Ahmed, Larry D. Gruppen, and Sally A. Santen. Pushing critical thinking skills with multiple-choice questions: Does bloom's taxonomy work?, 6 2018.

[45] Lishan Zhang and Kurt VanLehn. Evaluation of auto-generated distractors in multiple choice questions from a semantic network. *Interactive Learning Environments*, 29(6):1019–1036, 2021.

[46] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. 9 2023.

Appendix A

Prompting Strategies

A.1 Multiple Choice

This prompt is injected within the human template as q_type when multiple choice is the type of question chosen.

”’

You should create multiple choice question/s. To write effective multiple-choice questions that target high cognitive skills, you should consider the following guidelines:

1. Create a focused stem: The stem of the question should be concise and clearly state the problem or situation being addressed. It should provide enough context for learners to understand what is being asked. If learners are being asked to assess an unseen scenario, describe the scenario using concrete details rather than in terms of the learned concepts. Remember – the learners should be able to draw their own connections between learned concepts and concrete scenario.

2. Avoid extra information: Ensure that the stem does not include irrelevant or excessive information. This can distract learners the key concepts being assessed and may lead to confusion.

3. Use plausible distractors: options, including the correct answer and the distractors, should be plausible and similar in length to avoid giving away the answer. Distractors should represent common misconceptions or errors that learners might make and be attractive to learners who lack a complete understanding of the topic. Effective distractors can often use learned concepts from the text in incorrect ways that could be believable to someone with incomplete understanding of the topic.

4. Avoid grammatical clues: Ensure that there are no grammatical clues or inconsistencies that give away the correct answer. Learners should rely solely on their

understanding and application of the topic to select the correct response.

5. Avoid giving away clues with the tone of words: when formulating a distractor, avoid using words with a negative connotation or tone that could make the distractor unattractive to learners. For example, in a question about the way to address an employee how has performed a task with very little apparent care or effort, compare the following distractors: - 'Micromanage the employee's every move to ensure they put in more effort' – the connotation of the phrase 'micromanage… every move' is negative, which could make it too obvious that this answer is incorrect. - '"Provide detailed guidance and close monitoring to ensure the employee is putting in their best effort" – this rephrased version carries a similar meaning but uses a positive tone and is therefore much more plausible as a distractor.

6. Include clear instructions: Provide clear instructions regarding how many options learners should select, whether they should choose the best answer or all that apply. Ambiguous instructions can confuse learners and affect their performance.

7. Keep the options homogeneous but differentiated: Make sure that the options are similar in terms of style and grammar. This prevents learners from easily identifying the correct answer based on differences in language or formatting. At the same time, options should be well-differentiated in their meaning.

8. Use a variety of question formats: Incorporate different question formats, such as multiple correct options, matching items, or scenarios, to keep the assessment diverse and engaging.

9. Write each option in a separate line.

‘‘‘

A.2 Bloom's Levels: Learning Objectives

Depending on the Bloom's taxonomy level selected by the user, a different prompt was included as the learning_objective variable. The four levels considered in this paper are remembering, understanding, applying and analyzing. Below are the prompts that were included for each Bloom's level.

A.2.1 Applying

”””

The questions should assess the learning objective of Application - learners' ability

to apply knowledge (concepts, approaches, principles, techniques, skills) and use it to solve problems or perform tasks in a new scenario.

For each question you develop, do the following: “‘ 1. Assign a number to the question (starting from 1)

2. Identify one or more key elements (concepts/principles/approaches/techniques/skills) taught in the textbook, which could be effectively applied by learners to a new scenario in ways that allows assessing and promoting their understanding and ability to apply these elements to unseen contexts. If what is taught in the textbook cannot be effectively applied to a new scenario, say ”The knowledge taught in the textbook cannot be effectively applied to a new scenario.”

3. Construct a fairly detailed scenario, using 30 to 50 words, that invites learners to effectively apply to the scenario the taught elements you identified in step 1. The scenario should be new, offering learners an opportunity to apply their learning in a fresh way that goes beyond the examples given in the textbook. At the same time, the learned concepts from the textbook should be sufficient for allowing learners to engage with the scenario in a meaningful way.

4. Using the scenario you constructed in step 2, develop a question that assesses the learner’s ability to apply the learned concepts you identified in step 1.

5. Answer the question you developed in step 3.

6. Explain the answer to the learner in a way that deepens their understanding of their learned concepts and promotes their ability to successfully answer similar questions in the future.

”””

A.2.2 Remembering

”””

The questions should assess the learning objective of Remembering - to memorize, recognize, and recall information, facts, details, and terms.

”””

A.2.3 Understanding

”””

The questions should assess the learning objective of Understanding - learners’ ability to comprehend the meaning of concepts. Learners should interpret, demonstrate,

classify, summarize, infer, compare and/or explain key ideas from the text.

For each question, pick one of the strategies below for writing questions assessing understanding. Pick strategies based on their fit with the textbook section and vary your choices if appropriate. Strategies:

“‘ - Exemplify: Ask learners to identify or create an example that does not appear in the textbook and instantiates a concept that does appear in it.

- Restate: Ask learners to identify or create a definition of a concept that is defined or explained in the textbook, but stated in a way that is very different from that in the textbook.

- Classify: Ask learners classify an example that does not appear in the textbook according to classification(s) that do(es) appear in it. It can be especially effective to combine two orthogonal classifications from the textbook in one question; for example, combining the classifications of gas/liquid/solid and toxic/nontoxic, and ask learners to classify Carbon Monoxide (gas, toxic).

- Infer: Construct an example or a scenario that do not appear in the textbook and ask learners to make an inference about the example/scenario based on what is taught in the textbook. - Summarize: Outline an idea that is articulated in a longer form over a section of the textbook.

”””

A.2.4 Analyzing

”””

The questions should assess the learning objective of Analyzing - learners’ ability to break down information into its constituent parts and identify patterns, relationships, or connections among them.

For each question assessing Understanding, follow one of the strategies below. Pick strategies that would make the most effective questions for the textbook section and vary the strategies as much as possible. Creating questions using these strategies requires a preliminary analysis step, which is specified for each strategy. Strategies:

“‘

- Assumptions: Ask learners to identify the assumptions they need to make for a certain proposition, theory, or claim from the text to be valid. Preliminary analysis: What is a proposition/claim/theory in the text that requires certain assumptions? What are these assumptions?

- Commonality: Ask learners to identify a common theme in the text or common characteristics among the taught elements, which are not stated explicitly in the text. Preliminary analysis: What is the common theme or what are the common characteristics?

- Comparison: Ask learners to compare and contrast entities or entity parts from the text in new ways that are not stated explicitly in it. The comparison should be insightful and deepen learners' understanding. Preliminary analysis: What are the entities for comparison? What are the dimensions for comparison?

- Classification: Ask learners to classify entities or entity parts from the text in new ways that are not explicit in it. The classification should be insightful and deepen learners' understanding. Preliminary analysis: What are the entities to classify? What is the classification system?

- Solution: Ask learners to find a solution that does not appear in the text to a problem/puzzle/tension that does appear in the text. Preliminary analysis: What is the problem/puzzle/tension?

- Prediction: Describe a scenario that does not appear in the text but relates to it and ask learners to determine plausible and/or implausible causes or outcomes of that situation. Preliminary analysis: What is the situation? What is the cause or outcome?

‘‘‘

’’’’’’

A.3 Feedback Refinement

A.3.1 Feedback Template

This is the prompt that is concatenated with the system prompt when questions are to be refined. The user must input all details of the original question including its Bloom's level and the textbook section is was generated from. The learning_obejective and q_type variables are generated in the same manner described for the original prompt. Questions are rated by the user on a 1 to 5 scale and the meaning of each of these ratings are included in the following subsection and inserted into the rating_meaning variable. Inputted feedback is included in the to_improve variable.

’’’’’’

You have previously generated a question based on a textbook passage. You must revise this question based on the feedback given and requirements listed below and the

textbook section delimited by triple backticks. Note: learners may not have access to the textbook section, so avoid making references to it.

{q_type}

The question was generated based on the following textbook section: Textbook Section: “‘

{textbook_section}

“‘

The question was developed based on the following learning objective instructions: {learning_objective}.

This question was generated to assess the following learning goal:

{learning_goal}

The following is the question that you had previously created.

Question: {question}

Options: {options}

Correct Answer: {correct_answer}

Explanation: {explanation}

The following is feedback on the generated question and answer. The overall rating is {rating} out of 5, which means they are {rating_meaning}.

The question should be improved by following these instructions: ’{to_improve}’

Revise this question based on the feedback given and the requirements listed below. The revised question should be a multiple choice question with a refined question stem, options, and explanation.

{next_step}

”””

{format_instructions}

A.3.2 Rating Meanings

- 1: ”way off the mark”,
- 2: ”not very good”
- 3: ”just so so”
- 4: ”quite good, but needs minor improvement”
- 5: ”right on!”

Appendix B

Surveys

B.1 Experiment 1: Secondary Teacher Survey

The full questionnaire given to secondary school teachers within the first experiment.

- **What is your occupation?** [Open Response]
- **How long have you been working within STEM education?** [Open Response]
- **Is the question relevant to the respective excerpt?**
  A. Yes, the question is relevant to the excerpt
  B. The question is relevant but requires knowledge beyond what is included in the text section
  C. No, the question is not relevant
- **Does the question contain correct information?**
  A. Yes, all information in the question is correct
  B. Mostly, but there are minor inaccuracies
  C. No, not all information contained in the question is correct
- **Is the question grammatical and well formed?**
  A. Yes, the question is grammatical and well formed
  B. The question contains minor grammar or syntactic errors, but these errors don't interrupt understanding the question.

C. The question contains major grammar or syntactic errors that make the question more difficult to understand.

- **Does the MCQ provide enough information to arrive at an answer?**

  A. Yes it provides enough information

  B. Somewhat, but additional information would be useful for clarity

  C. No, it doesn't provide enough information

- **Is there a correct answer listed in the options and is the option marked "correct" actually correct?**

  A. Yes, there is a correct answer and it is marked as the correct answer

  B. There is a correct answer but it is not given as the correct answer"

  C. No, there is no correct answer

- **Does the question only contain one correct answer?**

  A. Yes the question contains only one correct answer

  B. No, there are no correct answers given

  C. No, there are multiple correct answers

- **Are the options distinct from each other, ensuring they are unique choices?**

  A. Yes, they are completely unique between each other

  B. Some choices are unique, some are too similar

  C. No, they are too similar, making them repeated choices

- **Is the MCQ free from obviously-wrong options? Are all distractors plausible answers?**

  A. Yes there are no obviously wrong options

  B. Yes however the options give away the correct answer

  C. No, there are obviously wrong option(s)

- **Do the wrong answers target misconceptions that students may have?**

  A. Yes, all wrong answer choices target common misconceptions

B. Some wrong answer choices target common misconceptions

C. No, the wrong answer choices do not target misconceptions

- **Assuming you were teaching the same material, would you use this question in your course?**

  A. Yes

  B. No

- **What Bloom's Taxonomy level does this question target? If more than one apply, choose the level that is the most relevant to the question?**

  A. Analyzing

  B. Applying

  C. Understanding

  D. Remembering

  The following is the learning goal the question was tasked with assessing:

  (Example) ME.21 Create an energy story for the reaction catalyzed by glyceraldehyde-3-phosphate dehydrogenase, that discusses specifically the coupling of a redox reaction to a phosphate transfer

- **Does correctly answering the question help in progressing towards the learning objective?**

  A. Yes, to a large extent

  B. Yes, to some extent

  C. No, it does not

- **How do you feel about using artificial intelligence as a teaching aid? Do you believe that having access to AI tools designed for teachers would help your teaching practices? Are there any additional thoughts on this topic that you would like to share?**[Open Response]

B.2 Experiment 2 - Round 1

The full survey given to university level evaluators within the second experiment.

- **Is the question relevant to the respective excerpt?**
  A. Yes, the question is relevant to the excerpt
  B. The question is relevant but requires knowledge beyond what is included in the text section
  C. No, the question is not relevant
- **Does the question contain correct information?**
  A. Yes, all information in the question is correct
  B. Mostly, but there are minor inaccuracies
  C. No, not all information contained in the question is correct
- **Is the question grammatical and well formed?**
  A. Yes, the question is grammatical and well formed
  B. The question contains minor grammar or syntactic errors, but these errors don't interrupt understanding the question.
  C. The question contains major grammar or syntactic errors that make the question more difficult to understand.
- **Does the MCQ provide enough information to arrive at an answer?**
  A. Yes it provides enough information
  B. Somewhat, but additional information would be useful for clarity
  C. No, it doesn't provide enough information
- **Is there a correct answer listed in the options and is the option marked "correct" actually correct?**
  A. Yes, there is a correct answer and it is marked as the correct answer
  B. There is a correct answer but it is not given as the correct answer"
  C. No, there is no correct answer
- **Does the question only contain one correct answer?**
  A. Yes the question contains only one correct answer
  B. No, there are no correct answers given

C. No, there are multiple correct answers

- **Are the options distinct from each other, ensuring they are unique choices?**

  A. Yes, they are completely unique between each other

  B. Some choices are unique, some are too similar

  C. No, they are too similar, making them repeated choices

- **Is the MCQ free from obviously-wrong options? Are all distractors plausible answers that target student misconceptions?**

  A. Yes there are no obviously wrong options and all distractors target common misconceptions

  B. Yes however the options give away the correct answer

  C. No, there are obviously wrong option(s) and some of the answers do not target common misconceptions

- **Assuming you were teaching the same material, would you use this question in your course?**

  A. Yes

  B. No

The following is the learning objective the question was tasked with assessing:

(Example) ME.21 Create an energy story for the reaction catalyzed by glyceraldehyde-3-phosphate dehydrogenase, that discusses specifically the coupling of a redox reaction to a phosphate transfer

- **Does correctly answering the question help in progressing towards the learning objective?**

  A. Yes, to a large extent

  B. Yes, to some extent

  C. No, it does not

- **What Bloom's Taxonomy level does this question target? If more than one apply, choose the level that is the most relevant to the question?**

  A. Analyzing

B. Applying

C. Understanding

D. Remembering

- **The question was generated to match this Bloom's level: [Understanding]. If your selected Bloom's level does not match, what could be added to make the question of the desired level?**[Open Response]
- **What can be changed to make the question be of a higher quality? Please provide directed feedback on what changes could be made for the question to be better.** [Open Response]
- **When would using this question be most appropriate?**

  A. Exam Question

  B. Home Practice Question

  C. Reading Comprehension

  D. Other [With open response]

- **Please explain why you selected the question context you did in the last question.**[Open Response]
- **Do you have any additional feedback about this question?**[Open Response]

B.3 Experiment 2 - Round 2

The full survey given to university level evaluators within the second experiment after revising questions.

- **Is the question relevant to the respective excerpt?**

  A. Yes, the question is relevant to the excerpt

  B. The question is relevant but requires knowledge beyond what is included in the text section

  C. No, the question is not relevant

- **Does the question contain correct information?**

A. Yes, all information in the question is correct

B. Mostly, but there are minor inaccuracies

C. No, not all information contained in the question is correct

- **Is the question grammatical and well formed?**

A. Yes, the question is grammatical and well formed

B. The question contains minor grammar or syntactic errors, but these errors don't interrupt understanding the question.

C. The question contains major grammar or syntactic errors that make the question more difficult to understand.

- **Does the MCQ provide enough information to arrive at an answer?**

A. Yes it provides enough information

B. Somewhat, but additional information would be useful for clarity

C. No, it doesn't provide enough information

- **Is there a correct answer listed in the options and is the option marked "correct" actually correct?**

A. Yes, there is a correct answer and it is marked as the correct answer

B. There is a correct answer but it is not given as the correct answer"

C. No, there is no correct answer

- **Does the question only contain one correct answer?**

A. Yes the question contains only one correct answer

B. No, there are no correct answers given

C. No, there are multiple correct answers

- **Are the options distinct from each other, ensuring they are unique choices?**

A. Yes, they are completely unique between each other

B. Some choices are unique, some are too similar

C. No, they are too similar, making them repeated choices

- **Is the MCQ free from obviously-wrong options? Are all distractors plausible answers that target student misconceptions?**

  A. Yes there are no obviously wrong options and all distractors target common misconceptions

  B. Yes however the options give away the correct answer

  C. No, there are obviously wrong option(s) and some of the answers do not target common misconceptions

- **Assuming you were teaching the same material, would you use this question in your course?**

  A. Yes

  B. No

  The following is the learning objective the question was tasked with assessing:

  (Example) ME.21 Create an energy story for the reaction catalyzed by glyceraldehyde-3-phosphate dehydrogenase, that discusses specifically the coupling of a redox reaction to a phosphate transfer

- **Does correctly answering the question help in progressing towards the learning objective?**

  A. Yes, to a large extent

  B. Yes, to some extent

  C. No, it does not

- **What Bloom's Taxonomy level does this question target? If more than one apply, choose the level that is the most relevant to the question?**

  A. Analyzing

  B. Applying

  C. Understanding

  D. Remembering

- **The question was generated to match this Bloom's level: [Understanding]. If your selected Bloom's level does not match, what could be added to make the question of the desired level?**[Open Response]

- **What can be changed to make the question be of a higher quality? Please provide directed feedback on what changes could be made for the question to be better.** [Open Response]
- **When would using this question be most appropriate?**
    A. Exam Question
    B. Home Practice Question
    C. Reading Comprehension
    D. Other [With open response]
- **Please explain why you selected the question context you did in the last question.**[Open Response]
- **Do you have any additional feedback about this question?**[Open Response]
- **The following is the original generated question. [Question before revision from round 1 is provided.] Does the revised form of this question address the feedback provided?**
    A. The revised question significantly addresses the feedback provided
    B. The revised question partially addresses the feedback provided
    C. The revised question has not addressed the feedback at all
- **Please explain how the question has addressed the feedback provided. What parts of the feedback did it incorporate and what parts were left out?**[Open Response]

Appendix C

Participants' information sheet and Consent Form

Below is the combined Participant Information Sheet and Consent Form that participant's were presented with before continuing with the evaluation survey.

Project Title:	Mulitple Choice Question Generation: Motivating LLMs using Bloom's Taxonomy
Principal investigator:	Kobi Gal
Researcher collecting data:	Megan Morris

This study was certified according to the Informatics Research Ethics Process, reference number 643533. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?
Student Researcher: Megan Morris MSc Computer Science Supervisors: Kobi Gal, Avi Segal

What is the purpose of the study?
The purpose of this study is to evaluate the effectiveness of large language models in generating multiple choice questions when given text from educational content as a reference. The generation of these multiple choice questions will also be informed by Bloom's Taxonomy.

Why have I been asked to take part?

To evaluate the effectiveness of large language models in generating multiple choice questions, we will use human evaluators. Specifically, we would like to have educators with experience teaching science concepts at the secondary level and above provide feedback on the questions generated.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the PI. We will stop using your data in any publications or presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

The responses that you provide in the question evaluation form will be collected, and analysed to understand the quality of questions generated. Filling out the questionnaire should take a total of 45 minutes to an hour.

Are there any risks associated with taking part?

There are no significant risks associated with participation.

Are there any benefits associated with taking part?

There are no direct benefits in taking part in this study.

What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of 2 years.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the

researcher/research team as listed above.
All electronic data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint) and all paper records will be stored in a locked filing cabinet in the PI's office. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?
The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.
For general information about how we use your data, go to: edin.ac/privacy-research

Who can I contact?
If you have any further questions about the study, please contact the lead researcher, Megan Morris at s2603885@ed.ac.uk.
If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.
If the research project changes in any way, an updated Participant Information Sheet will be made available on http://web.inf.ed.ac.uk/infweb/research/study-updates.

Consent
By proceeding with the study, I agree to all of the following statements:

- I have read and understood the above information.
- I understand that my participation is voluntary, and I can withdraw at any time.
- I consent to my anonymised data being used in academic publications and presentations.

- I allow my data to be used in future ethically approved research.