Angel: A New Generation Tool for Learning Material based Questions and Answers

Ariel Blobstein[*]
Ben-Gurion University
arielblo@post.bgu.ac.il

Daniel Izmaylov[*]
Ben-Gurion University
izmaylov@post.bgu.ac.il

Tal Yifal
Data Innovation Lab
tal@datainnovationlab.ai

Michal Levy
Tel Aviv University
michalyot@gmail.com

Avi Segal
Ben-Gurion University
avise@post.bgu.ac.il

Abstract

Creating high quality questions and answers for educational purposes continues to be a challenge for educators and publishers. Past attempts to address this through automatic generation have shown limited abilities to generate questions targeting high cognitive levels, control question complexity and difficulty, or create adequate question-answer pairs. We take first steps toward addressing these limitations by introducing a new approach, named Angel, informed by recent developments in Large Language Models and Generative AI. Relying on advanced prompting techniques, automatic curation, and the incorporation of educational theory into prompts, Angel focuses on generating question answer pairs of varied difficulty while targeting higher cognitive levels. Questions and answers are automatically generated based on a textbook extract, with Bloom Taxonomy serving as a guide to the creation of questions addressing a diverse set of learning objectives. Our experiments compare Angel to several baselines and demonstrate the potential of informed generative models to create high-quality question answer pairs that cover a diverse range of cognitive skills.

1 Introduction

Generating high-quality questions and answers is of paramount importance in the educational sector. Exam-style questions serve as an essential educational tool for assessment as well as a catalyst for student learning. They provide opportunities for students to practice retrieving information, focus attention on key learning materials, reinforce learning through repetition of core concepts, and motivate engagement in learning activities [21]. However, manually creating these questions is a complex task, requiring significant resources, training, and experience. Automatic questions and answer generation techniques have emerged as a potential solution to these challenges, offering the possibility of constructing high-quality questions efficiently and cost-effectively. However, until recently, these techniques have primarily employed a template-based methods. Past reviews highlight the limitations of these approaches, particularly their inability to generate questions that target high cognitive levels, control question complexity and difficulty, or provide constructive feedback to the learner [26]. Furthermore, these approaches have been constrained by their tendency to rely on the existence of a pre-formulated answer for the generated question [4, 5]. Recent advances in Generative AI, known for their exceptional text generation capabilities, offer still untapped potential for automatic question and answer generation [2].

[*] Equal contribution: These authors contributed equally to this work.

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

We propose a novel approach, Angel, a generative tool for learning material-based questions and answers. Angel uses generative AI and employs advanced prompting techniques, automatic curation, and the integration of educational theory to create high-quality questions and answers. The generation process is guided by textbook material, with Bloom's Taxonomy serving as a framework for creating questions addressing diverse learning objectives and cognitive levels [3, 7]. Our LLM and human based evaluations compare Angel to several baseline and ablation conditions. The results demonstrate the potential of generative models, when coupled with educational theory, to create high-quality, cognitively diverse questions and answers in educational settings.

2 Related Work

Our research relates to past work on AI-based question and answer generation, generative AI-based approaches for content selection and evaluation, and taxonomies for the classification of learning objectives. We elaborate on each one in turn.

QA Generation Past work has mostly focused on question generation given a context (text) and an existing answer [5, 4, 8]. Popular datasets for such tasks are the SQuAD [18] and NQ [9] datasets. Recently, generative models have been used to generate questions in Q&A settings, as well as evaluate questions generated by other models. Nguyen et al. [16] trained a GPT-3 model to evaluate questions generated by a T5 transformer-based model. Their results demonstrate high correlation between GPT's ratings and the consensus between the ratings of two human experts. Alberti et al. [1] introduced a novel method of generating synthetic question-answering corpora by combining models of question generation and answer extraction, and by filtering the results to ensure roundtrip consistency. They demonstrate state-of-the-art results on two datasets. This work is one of the few works which generated both questions and answers given a textual context and it forms one of the baselines in our paper. Finally, research in introductory computer science has shown generative LLMs to be effective at generating code and explanations of the code for entry-level programmers [19, 10] when using ChatGPT or variants specifically fine-tuned on large code datasets.

Generative AI Content Curation and Evaluation Recent works have developed state-of-the-art approaches for content generation, curation (selection), and evaluation using generative LLMs. Yuan et al. [25] have developed a generation and selection approach to improve the quality and diversity of generative LLMs stochastic output. Specifically, they propose two prompt-based approaches for selecting high-quality questions from a set of LLM-generated candidates. They empirically demonstrate the efficacy of their approach compared to greedy generation using automatic and human based evaluations. The selection approaches developed in this paper draw inspiration from this work. Li et al. [11] developed a self-augmentation and self-curation approach using the LLaMA [22] generative model to create high-quality content samples for automatic labeling of human written text. Self-curation is achieved by careful prompt engineering, instructing the model to rate the quality of a candidate pair on a 5-point scale. Their approach outperformed all other LLaMA-based models on the Alpaca leaderboard [12], demonstrating highly effective self-alignment. Liu et al. [14] developed G-EVAL, a framework for using large language models with chain-of-thoughts (CoT) and a form-filling paradigm, to assess the quality of natural language generation outputs. Their approach utilized an auto-chain-of-thought mechanism, which harnessed the LLM to generate its own instructions. Their results demonstrate that G-EVAL achieves a high correlation with human annotators on summarization tasks, outperforming all previous methods by a large margin. We draw inspiration from this approach when developing our own LLM-based evaluation methods and auto-chain-of-thought mechanisms.

Learning Objective Taxonomies The Bloom Taxonomy [3], proposed in 1956, is the leading taxonomy used in education to develop and assess the cognitive complexity of educational content. The revised taxonomy [7], developed by Bloom's partner and co-author, is a hierarchical list of thinking measuring six cognitive levels that students may be required to demonstrate when engaging with educational material. These skills move from the lowest cognitive level to the highest cognitive level and include (1) Remembering - recalling details (the lowest level), (2) Understanding - comprehending meaning, (3) Applying - applying knowledge, (4) Analyzing - identifying patterns, (5) Evaluating - making judgment and critique, and (6) Creating - generating new ideas (the highest level). The taxonomy has been used extensively in the past to evaluate and improve the different cognitive levels elicited by educational questions developed by teachers and publishers [24, 6, 27, 15]. In addition, the Bloom taxonomy was instrumental in demonstrating that most educational questions developed in the past were of the Remembering type and pointed to the need to develop material which requires higher cognitive skills such as Analyzing, Evaluating and Creating [7]. In this work, we use Bloom's Taxonomy to instruct the generative model during the question and answer creation process and to evaluate the outcomes of the compared approaches. Finally, some recent work has used GPT4 to develop Bloom Taxonomy-based course material in educational settings [20]. Contrary to that, we focus on generation questions and answers when the course material is available, and utilize the much cheaper GPT3 model.

3 Method

Angel is an LLM-based generator that takes educational text as input and produces questions and answers of different difficulty and cognitive complexity levels. It works in several stages, as explained below.

3.1 Questions and Answers Generation

Angel's first stage is a prompt-based generator that contains instructions for creating questions and answers in varying difficulty levels for the given education text. We augment the instruction with three informed additions: (1) A "Chain-of-Thoughts" (CoT) command with instructions for the LLM to offer a sequential thought process on how to respond to the question using the text, followed by the produced answer. CoT was shown in previous work to improve the quality of created content by LLMs [23]. (2) A few shot prompt extension which includes examples of questions taken from the educational text itself, if available. Past research has demonstrated the efficacy of a few shot-based approaches with generative models (e.g. [13]). We hypothesize that adding human-created questions from the textbook itself will enhance the LLM's generation capabilities. (3) A Bloom Taxonomy-based instruction that guides the LLM in creating questions and answers aligned with the revised Bloom Taxonomy - Cognitive Domain categories. This prompt also includes the definition of the 6 levels of Bloom's Cognitive Taxonomy. We hypothesize that this instruction will enrich the questions and answers generated, leading to more questions created at higher levels of the taxonomy. The generation stage results in multiple <paragraph, question, answer, difficulty-level> tuples with varying difficulty levels (easy, medium, high) for all the educational content paragraphs provided to Angel. The prompts used for the generation stage are specified in Appendix A.

3.2 Self Augmentation

We augment the created result-tuples by utilizing the LLM's ability to create different responses with high-temperature settings. By executing each instruction repeatedly with a temperature setting of 0.9, we end up with a collection consisting of multiple question and answer variations for each paragraph to be provided in the follow-up self curation step.

3.3 Q&A Self Curation

Sample curation deals with selecting high quality questions and answers from the created Q&A samples. Notably, recent work in generative models developed n-gram and round-trip consistency-based approaches for sample curation [25, 11]. These approaches rely on the idea that effective questions and answers should better align (syntactically) with the information provided by the context. Unfortunately, this approach may not be suitable in our case, since Angel encourages the creation of questions for higher cognitive complexities, such as the Bloom Taxonomy Creation and Evaluation levels. In order to encourage such levels, one must be willing to tolerate reduced syntactic consistency between questions (or answers) and the provided context. Thus, we take a different approach to sample curation which focuses on cognitive complexity maximization. As such, our selection method chooses among each candidate group (i.e. a group of questions offered for each unit of text) the question with the highest Bloom Taxonomy level as identified by the LLM model during the generation phase. We hypothesize that such an approach will increase the Bloom levels of questions while not hurting the overall syntactic quality of the generated questions nor the overall correctness of the generated answers.

3.4 Evaluation Metrics

Previous work has shown that LLM-based metrics outperform reference-based and reference-free metrics in terms of correlation with human quality judgments, for open-ended and creative natural language generation (NLG) tasks. This was especially the case when augmented with Chain-of-Thought approaches [14]. Based on these insights, we use 4 evaluation metrics to analyze Angel's outcomes and compare to several baselines.

Questions Quality Evaluation We use GPT-3.5 [2] for this LLM-based evaluation, prompting it to answer a set of 5 meta questions about the final questions generated for each paragraph by the Angel algorithm. The meta questions are inspired by previous work [25] and ask about different aspects of the question's quality including clarity, relatedness to context, importance and answerability. We have extended the response scale of the LLM from a scale of 3 in the original paper to a scale of 5 for better granularity, and added a chain-of-thought instruction, asking the LLM to explain its grading as recommended in recent literature [14]. All prompts used for evaluation are specified in Appendix B.

Answers Quality Evaluation To develop the evaluation prompt for this measure, we use Auto-CoT [14], where the LLM is first instructed to propose the detailed evaluation instructions for answers given the paragraph and the question. As in prior work, instead of manually designing the evaluation steps for this task we instruct the LLM to generate such evaluation steps by itself, asking for a step wise approach and detailed explanations. The resulting prompt offered by the LLM is used (after human approval) for the answer quality evaluation. In this evaluation step answers are inspected for completeness, relevance, clarity, coherence, conciseness, correctness, depth, tone and engagement and are automatically scored on a 1-5 rating scale.

Bloom's Taxonomy Learning Objectives The above metrics do not address education-specific evaluation, thus we develop an LLM-based evaluation scheme for analyzing the Bloom Taxonomy learning objective levels of each generated question. As mentioned earlier, we hypothesize that Angel's generation approach (which explicitly references the taxonomy and its definition in the generation instructions) will create higher-quality questions with better coverage of Bloom's higher Taxonomy levels compared to baselines. The prompt for this evaluation step instructs the model to judge each taxonomy level for each question and score it on a scale of 1-5. In addition, we provide the LLM a short definition of each one of Bloom's levels (Remembering, Understanding, Applying, Analyzing, Evaluating, Creating) as part of the prompt instruction in order to avoid ambiguity in the definition of each category. See Appendix B for the full prompt.

Human Evaluation Human evaluation consists of human ratings on a subset of the questions and answers generated by the different conditions compared in this research. We solicit human scoring from one educational professional. Paragraphs provided to the human evaluator were randomized and all condition information was anonymized. The human rater was asked to answer questions about the generated questions pedagogical soundness, the generated answers correctness and the main Bloom Taxonomy of every generated question. For the pedagogical soundness and correctness we asked the evaluator to rank it in a binary format (e.g. 1 will be pedagogically sound, while 0 is not). For the Bloom evaluation, the evaluator was asked to label the question according only one (main) Bloom Taxonomy category, where 1 corresponds to the lowest level (Remembering), 2 to the next level (Understanding), up to 6 which corresponds the the highest level (Creating). See Appendix B for the full questions to the human evaluator.

4 Experiments

Dataset For our experiments, we utilize an 8th-grade science education textbook in English from the Indian National Council of Educational Research and Training (NCERT) [3]. This textbook covers various science topics, such as microorganisms, crop production, coal and petroleum, plant and animal conservation, animal reproduction, force, pressure, and friction, with each chapter having 5-10 units. Our algorithms focus on generating questions for each unit separately, with our experiments

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://ncert.nic.in/textbook.php?hesc1=0-13

involving the first three chapters. Human evaluations are conducted on a sample of three units from the book.

Baselines We compare Angel to three alternative conditions. (1) T5 - a T5 [17] based approach for generating question and answer collection based on round-trip consistency [1] which demonstrated state of the art results in past research. (2) Simple - in this approach we simplify the generation stage and use only a simple instruction asking for question and answer generation, without requiring chain-of-thought, without supplying question examples from the book and with no Bloom Taxonomy based guidelines. Additionally, this condition does not contain any curation procedure. (3) AngelRC (Random Curation): The Angel algorithm without the question curation phase. In this approach, instead of the informed question curation approach we simply perform random curation.

Implementation Details In all experiments, we use the 3.5-turbo variant of OpenAI's GPT-3 (175B parameters) model. We use a temperature of 0.9 during the generation stage to facilitate model creativity, while setting the temperature to 0.2 in the evaluation stage to increase semantic equivalence while enabling some linguistic diversity. For the selection phase we instruct the model to create 3 alternatives for each unit and each difficulty level. GPT3.5-Turbo was accessed through OpenAI's provided paid API. The total cost of the API calls performed for this research was less than $20.

5 Results

Questions and Answers Quality We first compare all methods on questions and answers quality as judged by the LLM-based evaluators. Table 1 presents the generated question scores for the different conditions. Scores are given on a scale of 1 (lowest) to 5 (highest). As seen by the table, all conditions got high scores (greater that 4) on all question dimensions. Nonetheless, the Angel approach got the top scores in 3 out of the 4 inspected dimensions. We hypothesise that the lower score in answearability is due to the higher complexity level of the questions created by Angel. Additionally, we note that all generative based conditions (Simple, AngelRC and Angel) outperformed the T5 conditions in 3 of the 4 inspected dimensions. Table 2 presents the generated answer scores across conditions. Again we note that the generative based conditions outperform T5 in all measures. The Angel approach receives top scores in 5 out of the 8 inspected dimensions.

Table 1: Question Scoring

Condition	Clarity	Relatedness	Importance	Answerability
T5	4.63	4.93	4.70	4.52
Simple	**4.66**	4.96	4.81	**4.55**
AngelRC	4.51	4.98	4.82	4.29
Angel	**4.66**	**5.00**	**4.92**	4.42

Table 2: Answer Scoring

Condition	Relevance	Clarity	Coherence	Conciseness	Correctness	Depth	Tone	Engagement
T5	4.37	4.53	4.53	4.54	4.50	3.37	4.99	3.68
Simple	4.92	**4.63**	**4.66**	**4.77**	**5.00**	3.84	**5.00**	3.98
AngelRC	4.86	4.61	4.64	4.63	4.93	3.98	**5.00**	4.08
Angel	**4.93**	4.62	**4.66**	4.61	4.93	**4.05**	**5.00**	**4.09**

Bloom Taxonomy Levels of Created Questions Figure 1 presents the LLM-based Bloom Taxonomy scoring. As seen by the figure, all conditions demonstrate significantly higher scoring on the lower taxonomy levels, meaning that all conditions mostly generate questions with the Remembering and Understanding cognitive complexity levels. Nonetheless, Angel receives the top scoring for the higher Bloom levels, specifically for the Evaluating and Creating levels. This is an indication that steering the generative model with domain specific expertise (in this case about question cognitive complexity based on Bloom Taxonomy) can indeed result in the desired outcomes.

Human Evaluation Finally we present in figure 2 the results of the human rater for all conditions. The evaluator inspected 36 randomized pairs of questions and answers, 9 from each condition (while

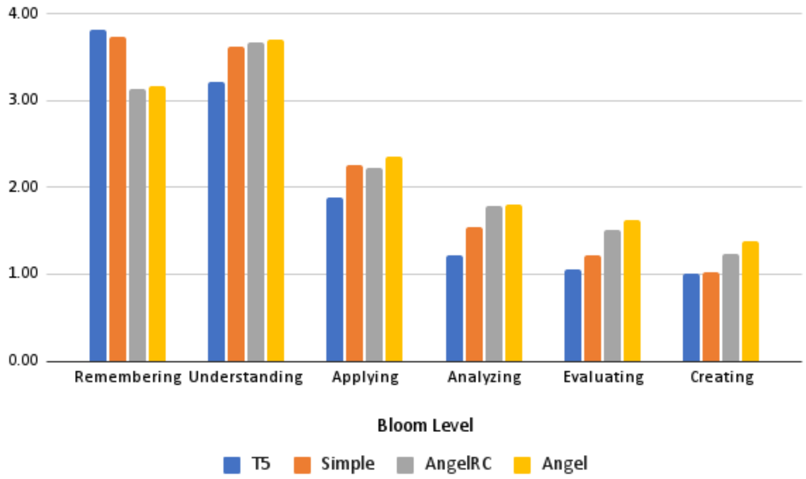

Figure 1: Bloom Taxonomy Scores

being blind to the condition). As seen in the figure, the evaluator judged all generated questions from all conditions as pedagogically sound, which correlates with the LLM-based scoring. This is by itself a strong finding for automated approaches to question generation. Additionally, the evaluator scores indicated that all generative based approaches outperformed T5 in answer rating and in Bloom Taxonomy level scoring, with AngelRS and Angel getting the highest scores for Bloom Taxonomy level scoring. We note that even for Angel, the human evaluator judged most questions as Remembering questions. Additional work and human scoring is needed to further analyze this outcome.

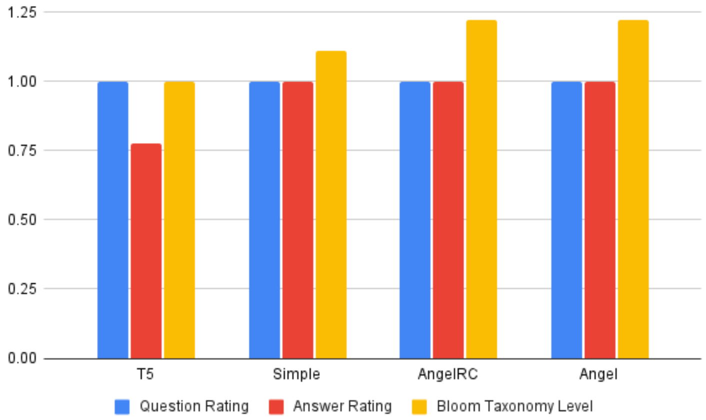

Figure 2: Human Evaluator. Q&A range is [0,1]; Bloom taxonomy range is [1,5].

6 Conclusion

In this study we investigate the practical problem of creating questions and answers for educational purposes given an educational textbook. We propose Angel, an informed prompting and curation based approach which builds on educational theory and question samples. Angel uses a generative LLM model for creating and curating high quality question answer pairs steering the creation process towards higher Bloom Taxonomy complexity levels. Our evaluations show that Angel outperformed other non-curated and non generative approaches on most of the inspected measures, and point to the potential of such generative endeavours in creating high quality questions and answers, especially when augmented with domain expertise. Future work should focus on increasing sample size to check for statistical significance, providing better human curated examples to the generative process to demonstrate higher cognitively complex questions, and using multiple LLMs for LLM-based scoring to alleviate for possible bias introduced by a single model.

References

[1] C. Alberti, D. Andor, E. Pitler, J. Devlin, and M. Collins. Synthetic qa corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416*, 2019.

[2] D. Baidoo-Anu and L. O. Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62, 2023.

[3] B. S. Bloom and D. R. Krathwohl. *Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain*. longman, 1956.

[4] X. Du, J. Shao, and C. Cardie. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*, 2017.

[5] M. Heilman and N. A. Smith. Good question! statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, 2010.

[6] E. S. Koç and T. Öntas. A comparative analysis of the 4th and 5th grade social studies curriculum according to revised bloom taxonomy. *Cypriot Journal of Educational Sciences*, 15(3):540–553, 2020.

[7] D. R. Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[8] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204, 2020.

[9] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[10] J. Leinonen, P. Denny, S. MacNeil, S. Sarsa, S. Bernstein, J. Kim, A. Tran, and A. Hellas. Comparing code explanations created by students and large language models. *arXiv preprint arXiv:2304.03938*, 2023.

[11] X. Li, P. Yu, C. Zhou, T. Schick, L. Zettlemoyer, O. Levy, J. Weston, and M. Lewis. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*, 2023.

[12] X. Li, T. Zhang, Y. Dubois, R. Taori, I. Gulrajani, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023.

[13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[14] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[15] T. Muhayimana, L. Kwizera, and M. R. Nyirahabimana. Using bloom's taxonomy to evaluate the cognitive levels of primary leaving english exam questions in rwandan schools. *Curriculum Perspectives*, 42(1):51–63, 2022.

[16] H. A. Nguyen, S. Bhat, S. Moore, N. Bier, and J. Stamper. Towards generalized methods for automatic question generation in educational domains. In *European conference on technology enhanced learning*, pages 272–284. Springer, 2022.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[18] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[19] S. Sarsa, P. Denny, A. Hellas, and J. Leinonen. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43, 2022.

[20] P. Sridhar, A. Doyle, A. Agarwal, C. Bogart, J. Savelka, and M. Sakr. Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives. *arXiv preprint arXiv:2306.17459*, 2023.

[21] W. Thalheimer. The learning benefits of questions. *Work Learning Research*, 2003.

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[24] A. C. Welch, S. C. Karpen, L. B. Cross, and B. N. LeBlanc. A multidisciplinary assessment of faculty accuracy and reliability with bloom's taxonomy. *Research & Practice in Assessment*, 12:96–105, 2017.

[25] X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, P. Lucas, H. Sauzéon, and P.-Y. Oudeyer. Selecting better samples from pre-trained llms: A case study on question generation. *arXiv preprint arXiv:2209.11000*, 2022.

[26] R. Zhang, J. Guo, L. Chen, Y. Fan, and X. Cheng. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43, 2021.

[27] S. L. Zorluoglu, A. Kizilaslan, and M. D. Yapucuoglu. The analysis of 9th grade chemistry curriculum and textbook according to revised bloom's taxonomy. *Cypriot Journal of Educational Sciences*, 15(1):9–20, 2020.

A Generation Prompt

A.1 Simple Generation Prompt

You are a middle school science teacher. You are given a paragraph from your textbook and you need to write final exam questions and answers for your students. You are supposed to write questions of 3 types: easy, medium, and hard.
The Paragraph: {{Paragraph}}

Figure 3: Simple Generation Prompt

A.2 ANGEL Generation Prompt

You are a middle school science teacher. You are given a paragraph from your textbook and you need to write final exam questions and answers for your students. You are supposed to write questions of 3 types: Easy, medium and hard.
Additionally, your generated questions should be based on the Bloom Taxonomy. Specifically, you should create more questions with higher Bloom Taxonomy learning objectives (applying, analyzing, evaluating, creating) as opposed to questions with lower learning objectives (reading, understanding).
For each generated question, you should specify its main Bloom Taxonomy learning objective and explain why you think this is the question's main Bloom Taxonomy objective.
Assume that the following are the Bloom Taxonomy learning objectives:
Remembering: This level involves recalling facts, details, or information. It is the most basic level of cognitive skill.
Understanding: At this level, learners comprehend the meaning of information and can explain it in their own words. They demonstrate comprehension of concepts and principles.
Applying: This level requires the application of knowledge and concepts to solve problems or perform tasks. It involves using information in new and different situations.
Analyzing: Analyzing involves breaking down information into its constituent parts and identifying patterns, relationships, or connections among them.
Evaluating: At this level, learners make judgments about the value or quality of ideas, solutions, or arguments. They can critique, assess, and defend their positions.
Creating: This is the highest level of cognitive skill, where learners can generate new ideas, concepts, or products. They can synthesize information from different sources to create something new.
You can see under "Examples:" below a few examples of past questions created for this paragraph. Please provide each answer with a step-by-step thinking for why the specific answer is the right answer.
Examples: {Examples}
The Paragraph: {Paragraph}

Figure 4: Angel Generation Prompt

B Evaluation Prompts

B.1 Question Quality Evaluation

You are an experienced educator for mid-school pupils. Your task is to receive questions written by a junior teacher for a given paragraph and to grade the questions based on 5 question that will be given to you. For each evaluation question there are five possible answers that should help you decide what your final score is. For each of the evaluation question you should write a score of 1-5 based on the answers.
You will be given a paragraph and questions that are being asked on this paragraph.
The Paragraph: {Paragraph}
The Questions for this paragraph: {{Question created for the paragraph}}

1. Is the question clear?
   1) It is not at all clear
   2) It is mostly unclear
   3) It is somewhat clear
   4) It is mostly clear
   5) It is very clear
2. Is the question related to the context of the attached document?
   1) It is not at all related
   2) It is mostly unrelated
   3) It is somewhat related
   4) It is mostly related
   5) It is closely related
3. Is the question asking about an important aspect of the context of the attached document?
   1) Not at all important
   2) Mostly unimportant
   3) Somewhat important
   4) mostly important
   5) It is very important
4. Can the question be answered using information in the attached document?
   1) No, answering the question requires completely different information
   2) Mostly not, answering the question requires a lot of additional information
   3) The question can be partially answered using information from the document
   4) The question can be mostly answered using information from the document
   5) The question can be perfectly answered using information from the document
5. What is your overall rating of the question generated based on the attached document?
   1) The question is very bad
   2) The question is quite bad
   3) The question is okay
   4) The question is quite good
   5) The question is very good

Figure 5: Questions Quality Evaluation Prompt

B.2 Answer Quality Evaluation

You are an experienced educator for mid school students. You are given a paragraph, a question that relates to the text of this paragraph and an answer for the given question. Your task is to evaluate if the answer is a good answer to the given question, considering the paragraph text. Evaluation steps:
1. Read the Paragraph: Start by carefully reading the paragraph provided. Understand the context, main points, and any relevant details.
2. Analyze the Question: Examine the question that relates to the paragraph. Ensure you have a clear understanding of what the question is asking for.
3. Review the Answer: Carefully read the answer provided and assess it based on the following criteria:

- Completeness: Does the answer address all aspects of the question, or is it missing key information?
- Relevance: Is the content of the answer relevant to the question, or does it contain irrelevant or off-topic information?
- Clarity: Is the answer written in a clear and understandable manner?
- Coherence: Is the response logically structured and organized, making it easy to follow?
- Conciseness: Is the answer concise and to the point, or does it contain unnecessary filler content?
- Correctness: Does the answer provide accurate information based on the paragraph text?
- Depth: Does the answer go beyond surface-level details and provide a comprehensive response to the question?
- Tone: Is the tone of the answer appropriate for an educational context, avoiding personal opinions and biases?
- Engagement: Is the answer engaging and interesting to the target audience (middle school pupils)?

4. Assign a Score: Use the 5-point scale to assign a score to the answer:

- Score 1: If the answer is incomplete, vague, off-topic, or controversial. If it contains missing content, promotional text, navigation text, or irrelevant information.
- Score 2: If the answer addresses the question to a minimal extent, providing only high-level details.
- Score 3: If the answer is helpful but lacks many details, contains personal experiences or opinions, or mentions external information.
- Score 4: If the answer is well-written, clear, and focused on addressing the question. It provides a complete and comprehensive response with minor room for improvement.
- Score 5: If the answer is a perfect response to the question. It's intentionally written, free of irrelevant content, of high quality, and demonstrates expert knowledge.

5. Document Scores: Keep a record of the scores and feedback for reference. This can be helpful for tracking progress and ensuring consistency in your evaluations.

6. Repeat for Each Answer: If you have multiple answers to evaluate, repeat the process for each one, ensuring a fair and consistent assessment.
The Paragraph: {{Paragraph}}
The Question for this paragraph: {{Question created for the paragraph}}
The Answer: {{Answer to the question}}

Figure 6: Answers Quality Evaluation Prompt

B.3 Bloom Taxonomy Level Evaluation

You are an experienced educator for mid school pupils. Your task is to receive questions written by a junior teacher for a given paragraph and to grade the questions based on the Bloom Taxonomy. For each bloom taxonomy learning objective you should give a score of 1-5 and an explanation of why you chose that score.
Use this scoring scale:

1- Minimal or No Coverage: A rating of "1" indicates that the educational question provides minimal or no coverage of the specified level of Bloom's Taxonomy. The question does not engage learners in thinking at that cognitive level.

2- Limited Coverage: A rating of "2" suggests that the question includes some elements or cues related to the specified Bloom's Taxonomy level, but it does not fully or effectively engage learners at that level.

3- Partial Coverage: A rating of "3" implies that the question partially addresses the specified Bloom's Taxonomy level. It involves some elements of thinking or skills associated with that level but may lack depth or complexity.

4- Adequate Coverage: A rating of "4" indicates that the educational question adequately covers the specified Bloom's Taxonomy level. It engages learners in thinking and tasks characteristic of that level, providing a reasonably challenging cognitive experience.

5- Comprehensive Coverage: A rating of "5" signifies that the question comprehensively and effectively covers the specified level of Bloom's Taxonomy. It engages learners in deep, complex thinking and problem-solving aligned with that level.

Assume that the following are the Bloom Taxonomy learning objectives:
Remembering: This level involves recalling facts, details, or information. It is the most basic level of cognitive skill.
Understanding: At this level, learners comprehend the meaning of information and can explain it in their own words. They demonstrate comprehension of concepts and principles
Applying: This level requires the application of knowledge and concepts to solve problems or perform tasks. It involves using information in new and different situations.
Analyzing: Analyzing involves breaking down information into its constituent parts and identifying patterns, relationships, or connections among them.
Evaluating: At this level, learners make judgments about the value or quality of ideas, solutions, or arguments. They can critique, assess, and defend their positions.
Creating: This is the highest level of cognitive skill, where learners can generate new ideas, concepts, or products. They can synthesize information from different sources to create something new.
The Paragraph: {{Paragraph}}
The Questions for this paragraph: {{Question created for the paragraph}}

Figure 7: Bloom Taxonomy Evaluation Prompt

B.4 Questions for Human Evaluator

1. Rate each question as either pedagogically sound (score=1) or not (score=0). A pedagogically sound question is one that pertains to the paragraph content and is intended to assess the domain knowledge of the student. A question is classified as not sound if it is vague, unclear, or not about assessing domain knowledge (similar to [16]).

2. Rate each answer as either correct given the paragraph text and the question (score=1) or incorrect (score=0).

3. For each question, identify its main Bloom Taxonomy Learning Objective. If the question covers several learning objectives, specify the main one. Assume that the following are the Bloom Taxonomy learning objectives: <Here we give the same definition of the Bloom Taxonomy given to the GPT model>.