Reducing Human Effort in Evaluating Small and Medium Language Models as Students and as Teachers

Oleh Prostakov
prostakov.pn@ucu.edu.ua
Ukrainian Catholic University
Lviv, Ukraine

Viacheslav Hodlevskyi
slavagodlevsky86@gmail.com
SET University
Kyiv, Ukraine

Nassim Bouarour
nassim.bouarour@univ-grenoble-alpes.fr
Univ. Grenoble Alpes
Saint Martin D'Hères, France

Adam Sanchez-Ayte
adam.sanchez@uness.fr
UNESS: Université Numérique en Santé et Sport
Saint Martin D'Hères, France

Noha Ibrahim
noha-ibrahim@univ-grenoble-alpes.fr
Grenoble Polytechnic Institute
Saint Martin D'Hères, France

Sihem Amer-Yahia
sihem.amer-yahia@univ-grenoble-alpes.fr
CNRS, Univ. Grenoble Alpes
Saint Martin D'Hères, France

ABSTRACT

Multiple Choice Questions (MCQs) are commonly used by teachers to assess student understanding, but generating high-quality MCQs is a demanding task. Large Language Models (LLMs) offer a potential solution, yet their use raises concerns about privacy, cost, and energy consumption, especially in educational settings. In this paper, we present a simple and reproducible evaluation framework designed to assess the ability of small and medium-sized LMs to answer (LM as student) and generate (LM as teacher) high-quality MCQs. The framework uses a set of clearly defined measures, such as syntactic correctness, relevance to source material, distractor quality, and answer consistency, to provide a detailed analysis of model performance. We applied the framework to evaluate several language models and found that each exhibits distinct strengths and weaknesses across different metrics. Notably, some small models—such as Phi-3.5-mini and Llama3.1:8b—outperform larger peers in specific areas, demonstrating that model size does not always correlate with overall quality. These findings empower teachers to choose models that best align with their goals and priorities, reinforcing their agency while highlighting the practical value of lightweight models in educational settings. We also outline future work, including targeted fine-tuning to improve model performance on specific MCQ quality dimensions.

VLDB Workshop Reference Format:
Oleh Prostakov, Viacheslav Hodlevskyi, Nassim Bouarour, Adam Sanchez-Ayte, Noha Ibrahim, and Sihem Amer-Yahia. Reducing Human Effort in Evaluating Small and Medium Language Models as Students and as Teachers. VLDB 2025 Workshop: DaSH: Data Science with Human in the Loop.

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/Its-OP/slm-mcq-eval.

Proceedings of the VLDB Endowment. ISSN 2150-8097.

1 INTRODUCTION

Multiple Choice Questions (MCQs) are widely used by teachers to assess students' understanding, progress, and analytical skills, as outlined in Bloom's taxonomy.[1] However, creating good MCQs takes time and effort, especially to ensure relevance, appropriate difficulty, and coverage of key concepts.

Challenges and Goal. Large Language Models (LLMs), both general and specialized, have been used to generate MCQs and can produce reasonably good results. Yet, they raise ethical concerns, such as exposing proprietary educational content, high energy consumption, and financial costs. These issues often discourage teachers from sharing their materials with such models.

Locally deployable small and medium-sized language models can address these concerns. But using them effectively is challenging: there is no framework that helps identify the model to choose, how to configure it (e.g., precision, temperature), or how to evaluate the quality of the questions it generates.

In this paper, we present such a framework with the goal of helping teachers generate MCQs directly from their own teaching material, on their own devices and infrastructures. Most importantly, our framework formalizes evaluation measures to assess the quality of generated MCQs, and provides a reproducible pipeline that helps teachers evaluate any language model on any dataset. We demonstrate its use in this paper by evaluating several small and medium-sized models on a specific educational dataset.

Contributions. Our aim is to develop a framework that reduces expert effort in generating high quality data, in our case, MCQs. Our first contribution is to define syntactic and semantic MCQ quality dimensions. Our framework generalizes existing MCQ quality measures [5, 9, 12]. Our measures are tested in the context of medical questions. Our second contribution is a reproducible methodology depicted in Figure 1 that takes any language model and returns a fine-grained evaluation of its ability to provide the correct answer to MCQs for which ground-truth answers are known (model as a student), and its ability to generate good MCQs from educational

[1] https://en.wikipedia.org/wiki/Bloom's_taxonomy

material (model as a teacher). Our code for the project is made available at our GitHub repository.

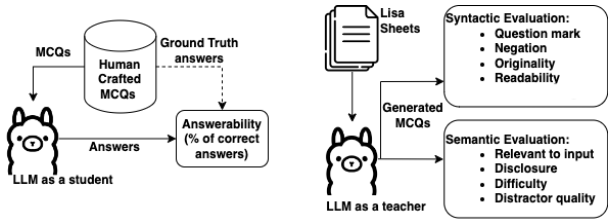

Figure 1: Our reproducible evaluation methodology for MCQ answerability and generation.

Findings. Our experiments revealed that LMs of different sizes have distinct strengths and weaknesses. While larger models generally perform better on ground-truth answers, small-sized models like Llama3.1:8b and Llama3.2:3b achieved similar results to larger models such as Mistral. Interestingly, smaller models like Phi-3.5:3.8b excelled in syntactic quality, sometimes outperforming bigger models. These findings show that performance isn't solely tied to model size, giving teachers the flexibility to choose models based on their specific needs. For instance, if the goal is to create simple questions to assess basic understanding, a model that scores well on syntactic accuracy and material relevance is preferable. If the focus is on testing deeper understanding, teachers may prioritize models that excel in distractor quality and semantic depth.

2 RELATED WORK

Several recent works relied on (most often large) language models to generate MCQs or to validate their outputs.

Meißner et al. [10] developed a quiz generation pipeline using GPT-4 and found that while it produced correct questions, its MCQs lacked originality. Olney [14] compared a fine-tuned Macaw model, instruction-tuned Bing Chat, and human-authored questions, concluding that LLMs perform almost as well as humans but fail differently. It was observed that Bing Chat sometimes omits correct answers, while Macaw generates non-distinct options. Three models, GPT-3.5, Llama3:70b, Mixtral:56b were checked for format, language, grammar, and relevance, and Llama3:70b was shown to perform best [11]. Scaria [12] evaluated five LLMs (Mistral, Llama, Palm, GPT*2) for question diversity and cognitive depth, using zero-shot, few-shot, and CoT prompting, as well as expert and LLM-based assessments. Hwang [5] generated MCQs in chemistry and biology using GPT-3.5, validated taxonomy alignment with RoBERTa, and plans to improve formatting with agent planning.

Most research focuses on generating high-quality, taxonomy-aligned (Bloom) questions that meet some, not all syntactic and semantic standards. Our framework captures and expands existing quality dimensions and tests a range of language models.

LLMs are also used to validate other LLM outputs. Shankar et al. [17] developed a platform that evaluates outputs based on LLM-generated criteria, aligning them with human classifications. Several studies [6] [2][3] proposed LLM-based evaluation measures, often outperforming traditional methods. Researchers have also explored automated correction of LLM outputs. Pan et al. [16] categorized correction methods into training-time, generation-time, and post-hoc approaches. Madaan et al. [8] introduced a self-refinement pipeline, while Fernandes et al. [4] reviewed enhancement techniques based on human feedback using Reinforcement Learning from Human Feedback. Subhankar [9] found strong consistency between questions generated with GPT-4 Turbo and human-assessed complexity, with GPT-4 aligning well with Bloom's taxonomy. However, only 13 out of 60 questions met high human-validation standards, mostly at the "Understanding" and "Remembering" levels of the taxomony.

3 MCQ QUALITY DIMENSIONS

To evaluate our models, we will first assess their ability to select correct answers based on ground-truth data, shedding light on their capabilities and the data they were trained on. We then evaluate their ability to generate MCQs using a set of rigorous quality measures—some drawn from prior work, others newly introduced to capture other key aspects. This section presents these measures. We will illustrate some of our measures using a running example. Consider the two example MCQs in Figures 2 and 3, both generated by GPT-4o-2024-08-06 from the same input material (Figure 8). Each MCQ includes a question and four answer options, with only one correct choice; the others are referred to as distractors. While both questions assess the understanding of depression, their syntactic and semantic dimensions differ, as described in more details in figures 4, 5, 6 and 7.

Which of the following best defines a major depressive episode (MDE)?
A) A condition primarily marked by elevated mood, excessive energy and bipolar disorder
B) A depressive episode that only occurs after a manic episode
C) A clinical state characterized by sad mood, anhedonia and low energy
D) A temporary emotional reaction to minor life stressors
Correct Answer: C

Figure 2: A good MCQ that is syntactically correct, does not disclose the correct answer in the question, is relevant to input material (depression), is of reasonable difficulty level, and has high quality distractors.

Not considered a major depressive episode any condition that does not include sad mood, anhedonia, low energy, and increased suicide risk, please choose the correct answer
A) Feeling a bit off sometimes
B) A medical issue with heart rate
C) A reaction to physical pain
D) A clinical state defined by sad mood, anhedonia, low energy, and higher risk of suicide
Correct Answer: D

Figure 3: A bad MCQ that is syntactically poor, has low quality distractors (not plausible answers), and that reveals the correct answer (as it is the longest and more detailed)

3.1 Syntactic dimensions

These dimensions focus on capturing the structure of an MCQ. We provide a measure for each dimension and set empirically-verified thresholds to return a Boolean value for each of them.

Question mark. We evaluate whether the question of an MCQ ends with a question mark ('?'). While not all MCQs must do so, we intentionally restrict ours to include a question mark.

Negation. To avoid confusion, we need to filter out questions that start with a negation by verifying if a question starts with one of a pre-defined set of negations[2] and assign 1 if it does, 0 otherwise.

Originality. Answers should avoid repeating the question, as this can lead to guessable answers. We propose to use trigram analysis and compare answer options to their questions. This returns 0 if more than 75% of the trigrams (from the question) also appear in an answer option. In that case the answer is considered too similar to the question.

A threshold of 50% is commonly used for trigram analysis. However, since the models were consistently scoring above 50%, we increased it to 75%.

Let $T(Q)$ and $T(A)$ be the sets of trigrams from the question and answer, respectively. The set of unique trigrams is:
$T_{unique} = T(Q) \setminus T(A)$.

The originality score O is then given by:

$$O = \frac{|T_{unique}|}{|T(Q)|} \quad \text{(with } O = 0 \text{ if } |T(Q)| = 0\text{)}$$

Readability. MCQs based on educational material must be written in a formal tone and rely on discipline-specific terms. We rely on the Flesch-Kincaid grading system[3] to define this measure. $R = 0.39\frac{total\_words}{total\_sentences} + 11.8\frac{total\_syllables}{total\_words} - 15.59$. According to common practice in Education, a value greater than 12 indicates readability at the college level.

Figures 4 and 5 illustrate examples of these syntactic quality measures using our running example.

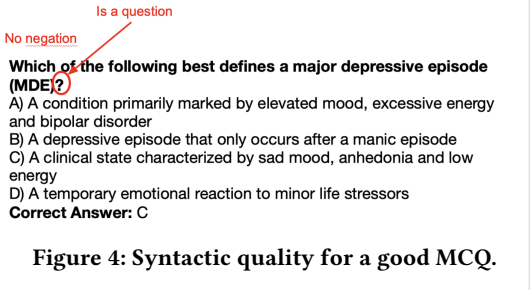

Figure 4: Syntactic quality for a good MCQ.

3.2 Semantic dimensions

These dimensions focus on capturing the content of an MCQ. As some of them are not easy to evaluate automatically, we follow prior work [18] and rely on a large LM as a judge.

Disclosure. LMs may generate MCQs in which the question discloses the correct answer. We prompt an LM judge (GPT 4o in our experiments) to evaluate that.

[2]https://inspe-sciedu.gricad-pages.univ-grenoble-alpes.fr/qcm/QCM_principes.html
[3]https://en.wikipedia.org/wiki/Flesch-Kincaid_readability_tests

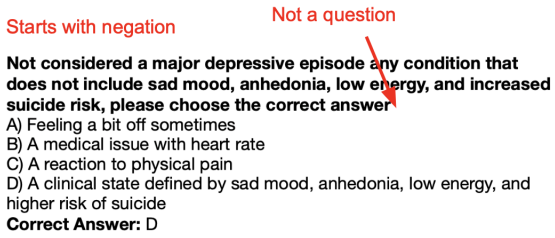

Figure 5: Syntactic quality for a bad MCQ

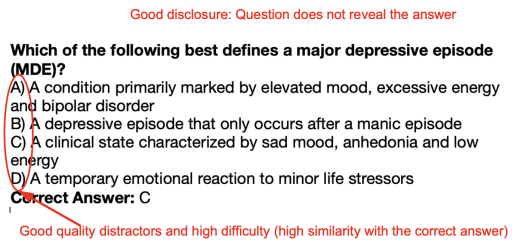

Figure 6: Semantic quality for a good MCQ

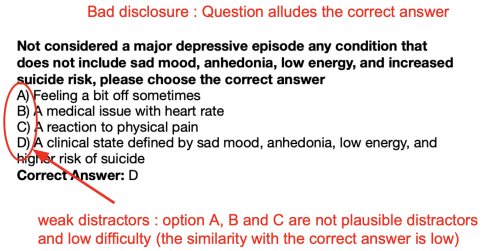

Figure 7: Semantic quality for a bad MCQ.

Relevance to input material. Since LMs are influenced by their training data, we check how relevant the generated MCQ is to the input educational material by measuring cosine similarity between their tokenized versions.

Difficulty. MCQ difficulty increases when distractors closely resemble the correct answer, which we measure using the average cosine similarity between their embeddings.

Gpt4-o answer alignment. GPT-4o answer alignment measures how often a model selects the same correct answer as GPT-4o, which was given the MCQs and the corresponding LISA sheet, reflecting its consistency with a strong reference model..

Quality of distractors. To capture how plausible and challenging distractors are, we use an LM judge to rate them from 1 to 5, and report the percentage of MCQs scoring 4 or higher

Figures 6 and 7 illustrate the semantic measures on our example

4 EVALUATION

We evaluated our framework from the perspective of a teacher aiming to compare several language models on a specific dataset and generate MCQs using their own material. Our evaluation focused on two main questions:

- **Model as a Student:** How well does a model perform when answering ground-truth questions?

- **Model as a Teacher:** Which models are most suitable for our teacher based on their desired quality dimensions?

Before presenting the results, we describe the selected models, the dataset, and the framework's infrastructure.

4.1 Infrastructure and models

For our comparative analysis, we established a controlled evaluation environment using OpenRouter's unified API system [4]. This infrastructure eliminates implementation variables while facilitating direct comparison under identical input conditions, ensuring that performance differences reflected genuine model capabilities rather than deployment inconsistencies.

We evaluated small models (Llama3.2:1b, Llama3.2:3b, Phi3.5:3.8b, and Llama3.1:8b) and medium sized models (Mistral3:24b, Llama3.3:70b) with $Q8\_0$ precision. We used GPT-4o (gpt-4o-2024-08-06) for quality dimensions that rely on an LLM as a judge. We vary the temperature values in $\{0.1, 0.5, 0.7\}$ and set the Top-p and context window to 1 and 4096 respectively. The model generates as many tokens as possible in JSON.

4.2 Dataset and ground truth

Knowledge Objective
ID: OIC-066-01-A
Parent Item: Diagnose: a depressive disorder, an anxiety disorder, obsessive-compulsive disorder, post-traumatic stress disorder, adjustment disorder, personality disorder
Level: A
Title: Understand the definition of a major depressive episode and recurrent depressive disorder
Description: None
Section: Definition
Contributors: Anonymized
Definition
A **major depressive episode (MDE)**, also known as "depression," is defined by a **low mood**, **anhedonia** (loss of interest or pleasure), **low energy**, along with affective, psychomotor, and physiological disturbances, and an **increased risk of suicide**. An MDE may occur as a single episode, as part of **recurrent depressive disorder** or **bipolar disorder**, or may be **comorbid** with another psychiatric disorder or a non-psychiatric medical condition. It can also follow **psychological and/or physical trauma**.
Recurrent depressive disorder is defined by the occurrence of at least one MDE **without a history of (hypo)mania**. The diagnosis is based on recurrence, severity, and clinical characteristics.

Figure 8: Example of a LISA sheet.

The UNESS[5] (Université Numérique en Santé et Sport) platform is designed for training medical students and professionals who face significant mental and academic demands during their studies. It comprises over 60k students of whom up to 4k are active simultaneously. The platform offers three types of exercises, aligned with Bloom's taxonomy and varying in difficulty levels: 29,952 Isolated Question Sequences (IQS), 4,757 Progressive Cases (PC) and 158 Critical Article Readings (CAR). In this work, we are interested in enriching the UNESS database with additional IQSs that correspond to MCQs on different medical specialties.

UNESS contains 4,500 LISA sheets, each of which can be seen as a Wikipedia document describing a medical specialty. Figure 8 provides an example of a LISA sheet. Each MCQ will be generated by giving one LISA sheet as input. The MCQ in Figure 2 and 3 were generated from this LISA sheet.

We use the MedMCQA dataset [15] to assess the ability of language models to answer multiple-choice questions with four options. It contains more than 194k high-quality AIIMS and NEET PG entrance exam MCQs covering 2.4k healthcare topics and 21 medical subjects. We randomly sample 200 MCQs per subject, by filtering out the 'Unknown' one, with MCQs having: (i) only one correct answer; (ii) uniform distribution across the correct options where an option is the correct one in 25% of the MCQs. This results in a subset of 4k MCQs, distributed over 20 subjects.

4.3 Prompt engineering

To assess disclosure and distractor quality, we carefully designed prompts for our judge model, GPT-4o. These prompts were structured to elicit precise, consistent, and interpretable outputs that align with the intended evaluation criteria.

To measure disclosure, we asked GPT-4o to determine whether a test-taker without relevant domain knowledge could guess the correct answer based solely on clues in the wording or structure of the question. The model was instructed to respond only with "True" or "False," minimizing ambiguity and simplifying result analysis.

To evaluate distractor quality, we prompted GPT-4o as shown below, to rate the plausibility of the incorrect answer choices on a scale from 1 to 5, based on how realistic or misleading they would appear to a test-taker.

Distractor quality prompt

You are tasked to evaluate the quality of the distractors of a multiple-choice question (incorrect options) on a scale of 1–5, where:
1 = POOR: Implausible, obviously incorrect, or unrelated to the question
2 = BELOW AVERAGE: Easy to eliminate, lacks plausibility for knowledgeable test-takers
3 = AVERAGE: Somewhat plausible but contains minor flaws that make it distinguishable
4 = GOOD: Plausible to most test-takers, represents common misconceptions
5 = EXCELLENT: Highly plausible, represents sophisticated misconceptions, requires deep understanding to eliminate
Provide only a numerical score from 1 to 5 that best represents the level of distractor quality.

Using a judge LLM is a practical alternative when no domain expert is available. However, in our future work, we plan to validate its outputs with a human expert to ensure accuracy and reliability.

4.4 Results

We report the results of evaluating language models on their ability to answer existing MCQs, as well as the syntactic and semantic quality of the MCQs they generate.

Model as a Student: Answerability. To evaluate the ability to answer existing MCQs, we prompted the models to answer the

[4] https://openrouter.ai/
[5] https://entrainement-ecn.uness.fr/

questions sampled from the MedMCQA dataset [15], normalized their answers, and then compared them with the ground truth answers provided by the same dataset. No additional material or contextual information were provided, requiring each model to rely solely on its pre-trained knowledge.

Ability to answer correctly (%)	**Temperature**		
Models	0,1	0,5	0,7
Llama3.2:1b	27	26,6	26,8
Llama3.2:3b	73,2	70,4	66,2
Phi-3.5:3.8b	58,3	58,4	57,7
Llama3.1:8b	74,4	57,15	46,05
Mistral3.2:24b	74	72,77	70,42
Llama3.3:70b	**92,1**	**92,38**	**91,97**

Table 1: Results on answerability when varying the temperature value with a precision model of q8_0.

As shown in Table 1, medium-sized models like Llama3.3:70b perform well in answering questions across different temperature settings. In contrast, smaller models show more variability in their results. Among them, Llama3.2:3b stands out for its consistency and performs surprisingly close to Mistral, despite having over six times fewer parameters. In contrast, the smallest model, Llama3.2:1b, often hallucinates and consistently selects option A as the correct answer, regardless of the question. Varying the temperature has little to no effect on some language models, while it has a drastic impact on others, such as Llama3.1:8b, so it may be an important factor to consider.

Model as a Teacher: MCQ Generation. Each model is instructed to generate MCQs with the same prompt and following the predefined JSON format:

```
{"question": {
   "question": "QUESTION_STATEMENT",
   "option_a": "OPTION_TEXT_A",
   "option_b": "OPTION_TEXT_B",
   "option_c": "OPTION_TEXT_C",
   "option_d": "OPTION_TEXT_D",
   "correct_option": "CORRECT_OPTION"
 }
}
```

We conducted MCQ generation for 3 different temperatures - 0.1, 0.5, and 0.7; we discovered that the models consistently performed best with temperature 0.1, showing slightly worse performance (within a couple of percents) at lower temperatures. We hence only report results with temperature 0.1; other results can be found in our repository.

To evaluate the ability to generate MCQs according to the expected valid JSON format, we prompted the models to generate one MCQ per LISA sheet from our 1,600 sheets. As shown in Figure 9, smaller models struggle to generate valid questions. "Not Generated" refers to generated content that is not readable. Given the same material, small models produced fewer valid MCQs and often required multiple runs to match the performance of larger models. Since the models did not complete the same number of MCQs, our evaluations are done on different sample sizes, we acknowledge that smaller test sets (e.g., 200 questions) may lead to less reliable estimates and higher variance in the results.

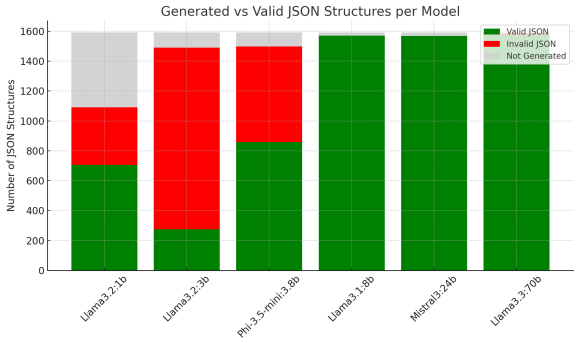

Figure 9: Number of valid generated MCQs per model for the same input material.

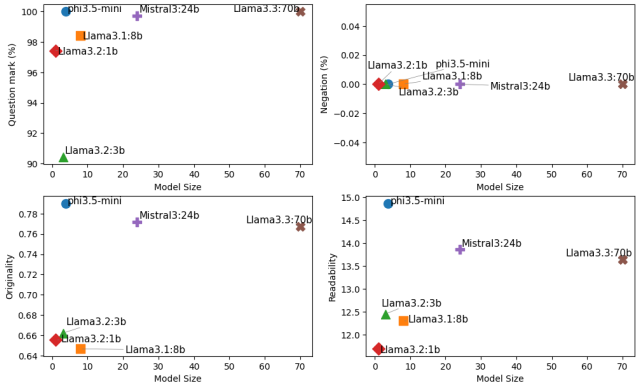

Figure 10: Results on syntactic dimensions (model size in billions of parameters).

Figure 10 reports performance on syntactic dimensions. Across originality, readability, use of question marks, and negation, Phi-3.5:3.8b outperforms not only the other SLMs but also the medium-sized ones. Its strong results are likely due to its extensive training process, which included supervised fine-tuning, reinforcement learning techniques, and preference-based optimization [1].

Figure 11 shows the performance of all language models on the semantic evaluation dimensions. In addition to these measures, we also evaluated how closely each answer selected by each model during question generation aligns with those provided by GPT-4o. Higher alignment scores indicate that the model's answers are more consistent with GPT-4o's judgments. To perform this comparison, GPT-4o was given the generated MCQs, the four answer options, and the corresponding LISA sheet material, and was asked to identify the correct answer.

Results vary by dimension: medium-sized models outperform small ones on disclosure, alignment with GPT-4o, and relevance, though Phi-3.5:3.8b remains competitive. Surprisingly, small and medium models both generate less plausible distractors, leading

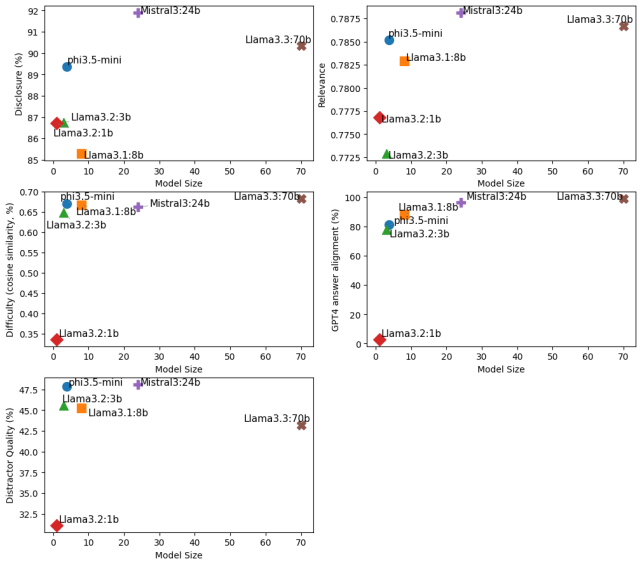

Figure 11: Results on semantic dimensions (model size in billions of parameters).

to significantly lower quality than GPT 4-o (87% of 'good' MCQs with threshold of 3). Once again Phi3-3.5:3.8b performs surprisingly well. Figure 11 also supports our earlier results about how well each LM answers MCQs. The 'GPT-4 answer alignment' plot shows that larger LMs perform better followed closely by Llama3.1:8b, while Llama3.2:1b tends to hallucinate and achieves very bad scores.

4.5 Limitations and Recommendations

This work has a few important limitations. First, while we used a powerful language model (GPT-4) as an automatic evaluator (LM as a judge), its outputs were not validated by a human expert. As a result, the reliability of some evaluation measures, particularly those requiring domain knowledge—may be limited. Second, the models we compared were not all tested on the same number of questions, which may affect the consistency and statistical reliability of the results. Future work should include expert validation of the LM judge's assessments and ensure uniform evaluation conditions across models.

By using our framework, teachers can assess how well a language model generates valid and well-structured questions, and how it performs according to our evaluation measures. This allows them to make informed decisions about which models to use and in which context. Our results show that medium-sized models are a strong option, but some smaller models, such as Llama3.1:8b and Phi-3.5-mini—also perform remarkably well.

That said, the performance of small models can vary across datasets, showing strengths in some areas and weaknesses in others. For example, Phi-3.5-mini produces high-quality syntactic MCQs, but often fails to generate properly formatted JSON, and about 20% of its questions do not align with GPT-4 in identifying the correct answer. This means that while Phi-3.5-mini can generate plausible questions and distractors acting as a good teacher, it may fail to select the right answer as a student.

As a result, teachers using such models should carefully review the generated questions instead of blindly relying on them. Alternatively, they can use our framework to test other models and find one that better fits their needs, thereby maintaining their agency. Indeed, our reproducible pipeline allows anyone to easily evaluate any newly released model.

5 CONCLUSION AND FUTURE WORK

We introduced a practical and reproducible framework that enables teachers to evaluate and generate MCQs using their own material and a variety of locally deployable language models. These models and our framework can be easily used through platforms such as LM Studio [7] or Ollama [13]. LM Studio and Ollama are user-friendly platforms that allow teachers to run language models locally on their computers, enabling them to test and evaluate models without needing advanced technical skills or internet access. Our approach addresses key concerns on privacy, LM deployment cost and energy consumption by focusing on small and medium-sized models that can run on personal devices.

Through systematic evaluation across syntactic and semantic measures, we showed that while larger models often perform well on semantic alignment and answer accuracy, some smaller models, such as Llama3.1:8b and Phi-3.5-mini, offer competitive and sometimes superior performance on syntactic quality. However, we also highlighted the variability and limitations of small models, including issues with output format and answer consistency.

Ultimately, our framework gives teachers tools to make informed choices about which LM best aligns with educational goals. While no single model is universally optimal, our results demonstrate that thoughtful evaluation enables targeted use of LMs in education, promoting teachers' agency and adaptability in question design.

In future work, we plan to fine-tune small language models to improve their performance on specific aspects of MCQ generation. Rather than aiming for general improvements across all dimensions, we will focus on targeted fine-tuning to address model-specific weaknesses. Another important aspect would be to fine-tune very small models to be able to generate valid JSON format, this could be done by using Low-Rank Adaptation (LoRA) specifically for the task of generating multiple-choice questions (MCQs) in JSON format based on provided context text.

More broadly, this targeted fine-tuning approach is a natural extension of our evaluation framework as it has the potential to make small models more robust and better aligned with educational needs, while maintaining the benefits of local deployment.

ACKNOWLEDGMENTS

This work was partially supported by DataGEMS, funded by the European Union's Horizon Europe Research and Innovation programme, under grant agreement No 101188416. This work was partially funded by the BPI France FRANCE 2030 call "Digital Commons for Generative Artificial Intelligence" (PARTAGES project DOS0245197). This research was supported in part by the NYU Center for Responsible AI, as part of the RAI for Ukraine program.

REFERENCES

[1] M. Abdin and all. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[2] Y. Chen, R. Wang, H. Jiang, S. Shi, and R. Xu. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. In *CoRR, abs/2304.00723*, 2023.

[3] D. C.-H. Chiang and H. yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 15607–15631. Association for Computational Linguistics*, 2023.

[4] P. Fernandes, A. Madaan, E. Liu, A. Farinhas, P. H. Martins, A. Bertsch, J. G. C. de Souza, S. Zhou, T. Wu, G. Neubig, and A. F. T. Martins. Bridging the gap: A survey on integrating (human) feedback for natural language generation. 2023.

[5] K. Hwang, S. Challagundla, M. M. Alomair, and L. K. Chen. Towards ai-assisted multiple choice question generation and quality evaluation at scale: Aligning with bloom's taxonomy. In *Generative AI for Education (GAIED): Advances, Opportunities, and Challenges. NeurIPS2024*, 2024.

[6] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Process- ing, EMNLP 2023, Singapore, December 6-10, 2023, pages 2511–2522*, 2023.

[7] LM Studio Team. Lm studio: Discover, download, and run local llms. https://lmstudio.ai/, 2025. Accessed: 2025-05-03.

[8] A. Madaan, N. Tandon, P. G. an d Skyler Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023*, 2023.

[9] S. Maity, A. Deroy, and S. Sarkar. How effective is gpt-4 turbo in generating school-level questions from textbooks based on bloom's revised taxonomy? In *Learnersourcing: Student-Generated Content @ Scale 2024*, 2024.

[10] N. Meißner, S. Speth, J. Kieslinger, and S. Becker. Evalquiz - llm-based automated generation of self-assessment quizzes in software engineering education. In *In Axel Schmolitzky and Stefan Klikovits, editors, Software Engineering im Unterricht der Hochschulen, SEUH 2024, Linz, Austria, February 29 - March 1, 2024, volume P-346 of LNI, pages 53–64.*, 2024.

[11] S. S. Mucciaccia, T. Meireles Paixão, F. Wall Mutz, C. Santos Badue, A. Ferreira de Souza, and T. Oliveira-Santos. Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2246–2260, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics.

[12] D. S. Nicy Scaria, Suma Dharani Chenna. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In *Artificial Intelligence in Education. AIED 2024*, 2024.

[13] Ollama Inc. Ollama: Run large language models locally. https://ollama.com/, 2025. Accessed: 2025-05-03.

[14] A. M. Olney. Generating multiple choice questions from a textbook: Llms match human per- formance on most metrics. In *Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023), Tokyo, Japan, July 7, 2023, volume 3487 of CEUR Workshop Proceedings, pages 111–128.*, 2023.

[15] A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.

[16] L. Pan, M. Saxon, D. N. Wenda Xu, X. Wang, and W. Y. Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. In *Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. CoRR, abs/2308.03188*, 2023.

[17] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. G. Parameswaran, and I. Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *CoRR, abs/2404.12272*, 2024.

[18] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.