Using LLMs to simulate students' responses to exam questions

Anonymous ACL submission

Abstract

Previous research showed that Large Language Models (LLMs) can be leveraged in numerous ways in the educational domain and, in this work, we study if they can be used to answer exam questions simulating students of different skill levels. From an educational perspective, this could enable to automatically evaluate learning and exam content and, from a computational linguistics perspective, it could help in understanding the learning process and knowledge of LLMs. By experimenting on three publicly available datasets, we show that it is indeed possible to prompt LLMs to simulate students of different skill levels using abstract scales, and share a prompt that proved effective in two different educational domains. We also show that, although the prompt generalises to different datasets, it does not generalise to different LLMs, and the LLMs do not seem capable to easily simulate students at specific levels of standardised educational scales.

1 Introduction

Large Language Models (LLMs) currently represent the state of the art in text generation, with some capable of generating human-like texts, such as OpenAI's GPT-4 (OpenAI, 2023), Llama 2 (Touvron et al., 2023) and Vicuna (Zheng et al., 2023). They are already being extensively used in a variety of domains, and in this work we focus on education, which could massively benefit from LLMs, as previous research discussed (Jeon and Lee, 2023; Kasneci et al., 2023; Caines et al., 2023). Specifically, we study whether it is possible to leverage LLMs to simulate the response patterns of students of different skill levels to exam questions. This could be leveraged for a variety of tasks, such as automatically evaluating learning content, estimating question difficulty, customising learning paths, and could also provide more insight into the learning process and the knowledge of LLMs. Previous research tried to simulate the responses of human participants to surveys with LLMs (Dillion et al., 2023; Argyle et al., 2023; Demszky et al., 2023; Aher et al., 2023), but nothing similar has been done for simulating students answering exam questions. There have been some concerns about the fairness of using LLMs instead of (or in addition to) human survey participants (Harding et al., 2023; Crockett and Messeri, 2023), and we agree that this is an important aspect to consider in the educational domain, as well. However, we believe that it might be less of an issue with respect to general-domain surveys, due to the factual nature of learning content and exam questions, which are built to evaluate domain knowledge and to minimise the effects that the wording has on the students' outcomes (Yaneva et al., 2019). In this work, we aim at answering the following Research Questions.

RQ1: can LLMs be prompted to answer Multiple Choice Questions (MCQs) while role-playing as (i.e., simulating) learners of different skill levels? Does this generalise to unseen data[1]?

RQ2: can LLMs simulate students at specific levels on standardised educational scales?

RQ3: How do these findings compare across different models?

Working primarily on GPT-3.5[2] and three publicly available datasets of science MCQs (*ARC*) and English reading comprehension MCQs (*RACE* and *CUP&A*), we show that it is indeed possible to prompt the LLM to answer exam questions with different levels of accuracy, and there is a positive correlation between the difficulty obtained from virtual pretesting with LLMs and the difficulty from pretesting with human learners. Also, for GPT-3.5 and our reference prompt, this behaviour is generalisable to previously unseen data (also from different educational domains) but, on the contrary, the effectiveness of the prompt does not generalise well to other LLMs. Lastly, even though we find that

[1] *Unseen* indicates data not used for prompt engineering.

[2] We use *gpt-3.5-turbo-0613*, except where explicitly said.

it is possible to prompt the models to role-play as students of different levels, it is not straightforward to simulate specific levels on standardised educational scales. The code, prompts, and LLM outputs are publicly available at *removed for anonymity, available in supplementary material for the review*.

2 Methodology

Search for the "best" prompt We work primarily with GPT-3.5, and prompt it to perform MCQ Answering (MCQA) simulating students of different skill levels. Crucially, in our setup, the LLM is shown only one question at a time, without having information about the other questions, nor the correctness of its previous responses. Similarly, the LLM is asked to simulate one student at a time, not to provide in a single response the answers of students of different levels. We work on three datasets, but perform prompt engineering only on one of them. Specifically, we do it on a *dev set* subsampled from *ARC* (science exams), and compare the model's behaviour when prompted with a variety of different prompts[3]. From this we get the "*reference prompt*", which is the one that leads to the best simulation of students' response patterns, according to the metrics defined in 3.2. Specifically, we are looking for increasing MCQA accuracy for increasing simulated levels, enough difference between the accuracy of low-skill and high-skill simulated students, and correlation between the results of virtual pretesting and pretesting with human learners. The reference prompt for *ARC* is shown in Table 1.

Analysis of the generalisation capabilities to unseen data We study the generalisation capabilities of the *reference prompt* as follows. We i) evaluate it on a different subset of data from *ARC*, and ii) we evaluate it on *RACE* and *CUP&A*, which contain English reading comprehension questions[4]. This approach might penalise the LLM, as the prompt was not engineered on these datasets, but we believe that it is a better way to study the generalisation capabilities of the proposed method.

Analysis of generalisation to other LLMs All previous steps are performed on *gpt-3.5-turbo-0613*, both prompt engineering to get to the *reference prompt* and evaluation of the generalisation capabilities. We also experiment on using the same reference prompts on different LLMs, to see whether the behaviour generalises. Specifically, we evaluate i) a different version of GPT-3.5 (*gpt-3.5-turbo-1106*), and ii) GPT-4 (*gpt-4-1106-preview*).

Analysis on standardised educational scales To understand if GPT-3.5 has *knowledge* of standardised educational scales and can simulate individuals at specific levels on such scales, we experiment with some minor variations of the reference prompt. Specifically, we experiment with: i) three language proficiency scales (*CEFR*, *IELTS*, and *TOEFL*), and ii) *exam marks* (A, B, C, D, F) [5]. We minimise the number of variables that might affect the model's output by starting from the reference prompt and performing the fewest modifications to add the information about the scales (e.g., from *"a student of level {x}"* to *"a student of CEFR level {x}"*)[6].

3 Experimental Setup

3.1 Experimental datasets

We experiment with three public datasets.

ARC, AI2's Reasoning Challenge dataset (Clark et al., 2018), is a MCQA dataset of questions from science exams. Each question is assigned a *grade* (from 3 to 9), which indicates the school grade that the question was built for. Although this is not a direct indication of question difficulty, questions with higher grades are meant for more advanced learners, and the *grade* has been used as a proxy for question difficulty in previous research (Benedetto, 2023). We work on a subsampled portion of the dataset: we use 350 questions as *dev set* and other 350 as *test set*. Both sets are sampled from the original *test* split with stratified sampling in order to have in both groups 50 questions for each grade.

RACE is a MCQA dataset of questions from English reading comprehension exams. We work on the version obtained by merging the original RACE (Lai et al., 2017) with RACE-c (Liang et al., 2019). Each question in the dataset is assigned one of three *levels* (*middle*, *high*, *college*), which indicates

[3]We only use zero-shot prompts and temperature=0.

[4]The prompt is actually slightly changed, swapping *a science exam* with *an English reading comprehension exam* and adding the text of the reading passage, to reflect the different nature of these datasets. The rest of the prompt is untouched.

[5]We also consider iii) *school grades* and iv) a non-standardised scale (*[beginner, intermediate, advanced]*). These are not the core of the paper and we briefly show the results in Appendix B.3 and B.4.

[6]We are aware that a negative result in this experiment does not necessarily prove that the LLM is not capable of representing these scales at all, but we argue that it is still valuable as it would suggest that even if it might be possible to use them, it is certainly not straightforward to do so.

Table 1: Reference prompt for the *ARC* dataset, the variable {X} in the system message is substituted with *one*, *two*, ..., *five* to indicate one of five student levels. In this work, we use only the *index* from the response, not the *question level* or *answer explanation*, but these are helpful to reach the desired behaviour in the simulations. The reference prompt for *RACE* and *CUP&A* is the same, except two changes: i) *a science exam* is swapped with *an English reading comprehension exam* and ii) we add *Reading passage: "{passage}"* to the user prompt (before *Question*).

SYSTEM:
You will be shown a multiple choice question from a *science exam*, and the questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of level {X} would pick.
Provide only a JSON file with the following structure: {"question level": "difficulty level of the question", "answer explanation": "the list of steps that the students of level {X} would follow to select the answer, including the misconceptions that might cause them to make mistakes", "index": "integer index of the answer chosen by a student of level {X}"}
USER:
Question: "{question}"
Options: "{answer options}"

the school level of the target students. Similarly to *ARC*, although this is not a direct indication of question difficulty, it has been used as a proxy for it in previous research, *middle* being the lowest difficulty and *college* the highest (e.g., by Loginova et al. (2021)). We work on a reduced set of 150 questions, obtained with stratified sampling from the *test* split, keeping 50 questions per level.

CUP&A[7] (Mullooly et al., 2023), is a MCQA dataset of questions from English reading comprehension exams. It contains questions aimed at students of different CEFR levels (from B1 to C2); it is not split into train, dev, and test. Similarly to the other datasets, we work on a stratified version, which is built by sampling 50 questions for each CEFR level, for a total of 200 questions. An important feature of this dataset is that, differently from *ARC* and *RACE*, it provides for all the questions an indication of the actual question difficulty, obtained from pretesting with real learners. This can be compared with the difficulty obtained from virtual pretesting performed with role-playing LLMs.

3.2 Evaluation metrics

Evaluating whether the LLMs are capable of simulating students' is not straightforward, especially considering the publicly available datasets. Indeed, the ideal evaluation would be to compare the response pattern of the LLMs with the response patterns of human learners, which is not available. As an alternative, we study each prompt by evaluating the responses for each simulated level as follows.

i) We study the MCQA accuracy of the LLMs when representing students of different levels; ideally, we want a monotonically increasing accuracy (i.e., higher role-played levels are more accurate).

ii) We study the MCQA accuracy, for the different role-played levels, on questions of different difficulty, to check whether lower role-played levels actually make mistakes on more difficult questions. This metric is partially hindered (for *ARC* and *RACE*) by the proxy used for the difficulty.

iii) For *CUP&A* we perform *virtual pretesting* using the responses from the LLM and compare the difficulty obtained from this with the reference value obtained from pretesting with real students.

4 Results and Analysis

4.1 Analysis of the reference prompt

Our first step consists in looking for the *reference prompt*, and this is done by iterating over a number of different prompts on the *dev* set of *ARC*, until we reach one with a satisfactory behaviour. Figure 1 shows how the MCQA accuracy of the LLM changes depending on the role-played level for different prompts; all the prompts shown here use non-standardised students' levels from *one* to *five*. For readability, we show only the reference prompt – which is the one that we selected as best performing on the *dev* set – and four other prompts which were explored at this stage. Compared to the

[7] The *Cambridge MCQs Reading Dataset* from Cambridge University Press & Assessment: https://englishlanguageitutoring.com/datasets/cambridge-multiple-choice-questions-reading-dataset

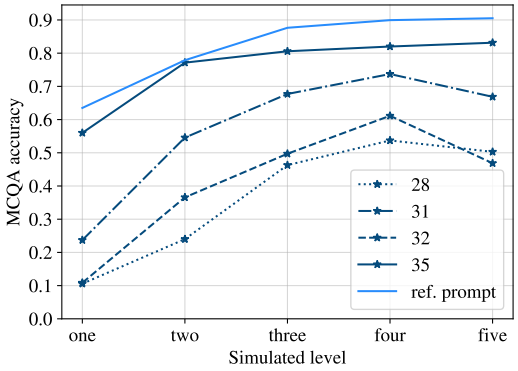

Figure 1: Comparison of the MCQA accuracy of GPT-3.5 on the *dev* split of *ARC*, when prompted with different prompts to simulate students of different levels. The prompts are available in Appendix A.1.

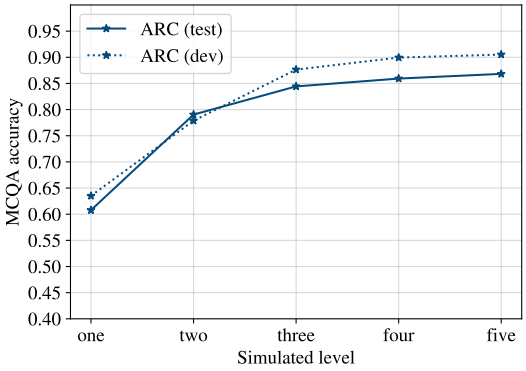

Figure 2: Comparison of the MCQA accuracy of GPT-3.5 on the *dev* and *test* splits of *ARC*, when using the *reference prompt* to simulate students of different levels.

reference prompt, i) prompt 28 adds a description about the meaning of students' levels; ii) prompt 31 removes the *answer explanation* and adds the text of the chosen answer; iii) prompt 32 adds the description of students' levels to prompt 31; and iv) prompt 35, the most similar to the reference prompt, renames the field *answer explanation* into *motivation*. They are shown in Appendix A.1.

A common issue is to have the highest MCQA accuracy for intermediate (simulated) levels – shown in the figure by prompts 28, 31, and 32 – and we observed this across a variety of different prompts, often triggered by minor changes.

Also, although the differences between the prompts shown in the figure are minor, the MCQA accuracy varies a lot (from 10% to 90%). Considering these prompts, we can easily say that prompts 28, 31, and 32 lead to a MCQA accuracy which is too low (the random baseline, without considering that students may guess the answer, is 25%).

Prompt 35 is close to the desired behaviour (the trend is monotonic), but it shows a significant step in accuracy between simulated levels *one* and *two*, and then the accuracy almost reaches a plateau, which is undesirable. The difference between this prompt and the reference prompt is only a renamed field in the output JSON required from the model, showing that even minor differences in the prompt can lead to relevant differences in the output.

4.2 Generalisation to unseen data

4.2.1 Analysis of MCQA accuracy

Figure 2 shows the behaviour of GPT-3.5 with the reference prompt on the *dev* and *test* portions of *ARC*. The behaviour is similar, with a monotonically increasing accuracy for increasing levels but, as expected, slightly worse on the test set.

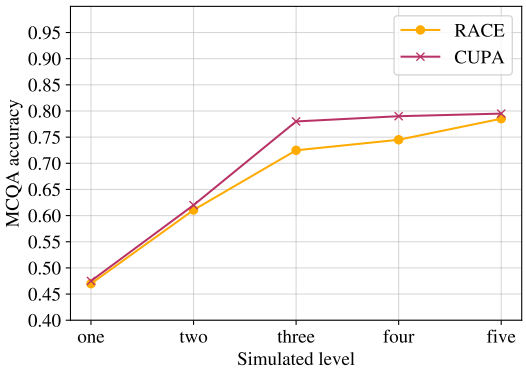

Figure 3: Evaluation of the MCQA accuracy of GPT-3.5 on *RACE* and *CUP&A*, when prompted with the *reference prompt* to simulate students of different levels.

We show in Figure 3 the evaluation of the reference prompt[8] on the two English reading comprehension datasets (*RACE* and *CUP&A*). The figure shows that, although these datasets were never seen while performing prompt engineering and they come from a different educational domain, the ability of the model to simulate students of different levels, when prompted with the reference prompt, transfers fairly well to them. Indeed, although for *CUP&A* the difference in accuracy for the three highest levels is very limited, for both datasets the MCQA accuracy is monotonically increasing.

Diving deeper, we plot in Figure 4 the same analysis for *RACE* but focusing separately on different questions *levels*. The figure shows that the trend

[8]Modified as described in Table1.

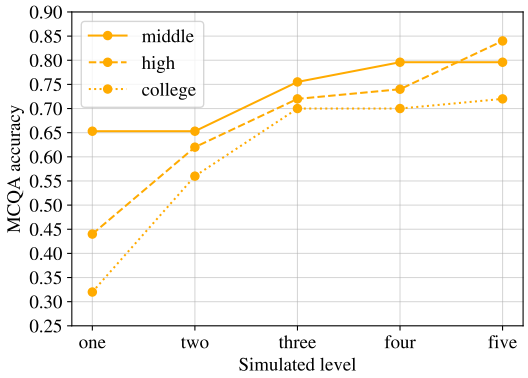

Figure 4: Evaluation of the MCQA accuracy of GPT-3.5 on *RACE* when simulating students of different levels, separately on questions of different *levels*.

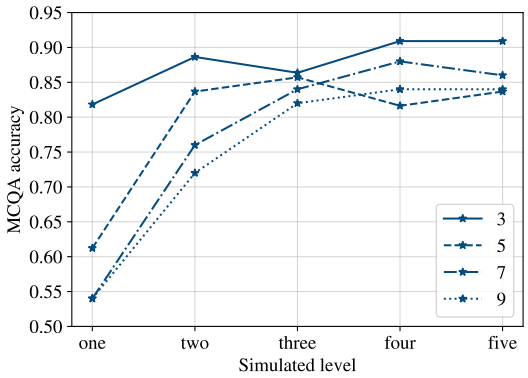

Figure 5: Evaluation of the MCQA accuracy of GPT-3.5 on *ARC* when simulating students of different levels, separately on questions of different *grades*.

of increasing MCQA accuracy for increasing simulated levels is visible across question levels. Also, if we look at the accuracy of a role-played level on questions of increasing levels, we can see that it consistently decreases, with the only exception of student level *five* on *high* questions.

Figure 5 shows the same analysis, but on *ARC*; we show only the odd *grades* to improve readability (the even *grades* are shown in Appendix B.1). The results are not as clean as on *RACE*: indeed, although we can see a general trend of increasing accuracy for increasing role-played levels, the trend is monotonic only for grade 9; grades 3 and 7 have one "drop" that affects monotonicity (level *three* and *five*, respectively), while grade 5 has several oscillations. Even though it is not always true that the same role-played level has lower accuracy on questions of higher grades, this trend is mostly visible for all grades, except grade 5 which seems to be the most problematic. This might also be due to the specific types of questions in *ARC*: indeed, even though most of the questions are knowledge questions for which it makes sense to define the difficulty, we observed that some do not necessarily get more difficult for higher grades (e.g., questions about safety equipment in the lab).

4.3 Virtual pretesting with role-playing LLMs

CUP&A provides for each question a quantitative measurement of difficulty obtained from pretesting with human learners. This enables us to evaluate the role-playing capabilities of the LLM by performing virtual pretesting and comparing the difficulty obtained from it with the reference value[9]. The results of the virtual pretesting are shown in Figure 6, which shows the correlation between the difficulty from the dataset (horizontal axis) and the difficulty obtained from virtual pretesting with the model simulating different student levels (vertical axis); ideally, we would want a perfect correlation between the two variables. It is worth mentioning that the two variables are on different scales (the "true" difficulty in $[30; 110]$ while the difficulty from virtual pretesting in $[0; 1]$) but this is not an issue: indeed, the difficulty from virtual pretesting could be converted to the other format with scale linking (Muraki et al., 2000). Even though the distribution of the dots on the plots is not really self-explanatory, the linear interpolation (the dotted line) shows that there is a positive correlation between the two variables. Specifically, the correlation coefficient[10] between the two variables is 0.13 ($pvalue = 0.06$), while a random baseline[11] leads to a correlation coefficient of -0.03 ($pvalue = 0.62$). To put the observed correlation in context, we also performed a brief Item Response Theory (IRT) simulation (Hambleton et al., 1991). This consists in simulating the responses of five "fake" students of prescribed skill levels to the questions of known difficulty from *CUP&A*, and perform pretesting with such responses. We consider students' skills equally spaced in the skill range, which is an ideal scenario and can be seen as an upper bound. This simulation led to a correlation of 0.43 ($pvalue = 10e^{-10}$).

[9] A short premise: at this stage, we are performing virtual pretesting with only five simulated students (GPT-3.5 role-playing as five students of different levels), which would be quite a small pretesting sample even with human learners.

[10] Computed with `scipy.stats.linregress`.

[11] It randomly assigns to each question a difficulty in the set `{0.0 ,0.2, 0.4, 0.6, 0.8, 1.0}`.

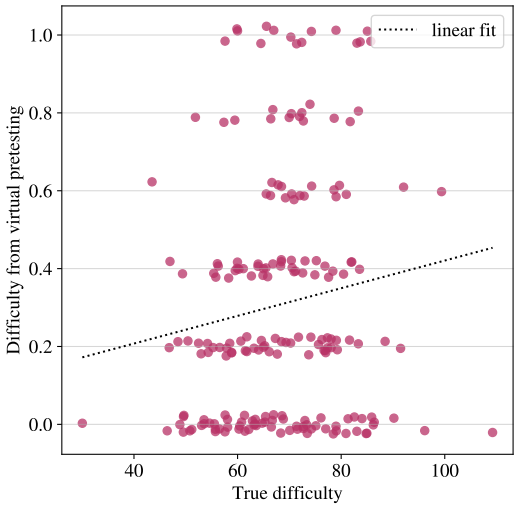

Figure 6: Results of the virtual pretesting on *CUP&A*. Each point represents a question, the position on the x-axis is determined by its "true" difficulty and the position on the y-axis is determined by the difficulty obtained from virtual pretesting ($\pm$ some random noise).

4.4 Generalisation to other LLMs

4.4.1 Newer GPT-3.5 version

Figure 7 compares the behaviour of the reference prompts when used on the latest (at the time of writing) version of GPT-3.5 (*gpt-3.5-turbo-1106*) and *gpt-3.5-turbo-0613*, which is the version used for prompt engineering and all the other experiments. The updated version of GPT-3.5 shows a similar behaviour, but there are some differences which arguably make it worse overall: indeed, for both *ARC* and *RACE* the highest MCQA accuracy is not obtained for level *five* but instead for level *three* and *four* respectively, and the model reaches a plateau at level *three* for both datasets. The behaviour on *CUP&A* is different, though: the newer version performs better in a way since it does not reach a plateau, but the difference in accuracy between the lowest and highest levels is smaller, which is undesirable. This last point is actually true across the three datasets: indeed, in almost all cases the newer *gpt-3.5-turbo-1106* leads to higher MCQA accuracy, and a narrower range of skill levels for virtual pretesting. These results suggest that prompts engineered for a specific version of GPT-3.5 should only be used on that specific version, as they might work differently when used on different versions.

4.4.2 GPT-4

We also study whether the behaviour is different when using the reference prompts to prompt the latest GPT-4 model (*gpt-4-1106-preview*), which outperforms GPT-3.5 in a variety of tasks. Figure 8 displays the behaviour of GPT-4 on the three datasets, and compares it with GPT-3.5. We can see that there is a monotonic trend of increasing MCQA accuracy towards higher simulated levels (except level *five* for *CUP&A*), but the accuracy of the lowest level is too high to be used for virtual pretesting (above 85% for all datasets). Again, we can see that the reference prompts are really effective only on the model used for prompt engineering. We find this particularly relevant since it shows that, even though it might be possible to perform prompt engineering on GPT-4 to reach a desired behaviour, it is not effective to engineer the prompts on GPT-3.5 and use them on GPT-4. Also, it shows that even though GPT-4 is a very powerful model, it is not immediate to get the desired behaviour from it, and might suffer of the *curse of hyper-accuracy*, also mentioned by Aher et al. (2023).

4.5 Evaluation on educational scales

To investigate the ability of GPT-3.5 to simulate students at specific levels of standard scales, we consider language proficiency scales (CEFR, TOEFL, IELTS), and exam marks (i.e., A, B, ..., F). Due to the different nature of the datasets and educational scales, we study the language proficiency scales on *RACE* and *CUP&A* only. As anticipated in Section 2, to better understand the contribution (positive or negative) of the educational scales on the model's behaviour, we minimise the differences with respect to the reference prompts. Specifically, we change the prompt only in the description of the student levels, without making other changes.

In addition to the experiments shown here, we also experimented with i) school grades and ii) another qualitative scale (*beginner, intermediate, advanced*). Since they are not the core of this work, we present the results in Appendix B.3 and B.4.

4.5.1 Language proficiency scales

To try and understand whether GPT-3.5 has some *knowledge* of common language proficiency scales, we modify the reference prompt to include levels from one of three scales: CEFR levels[12], IELTS

[12] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

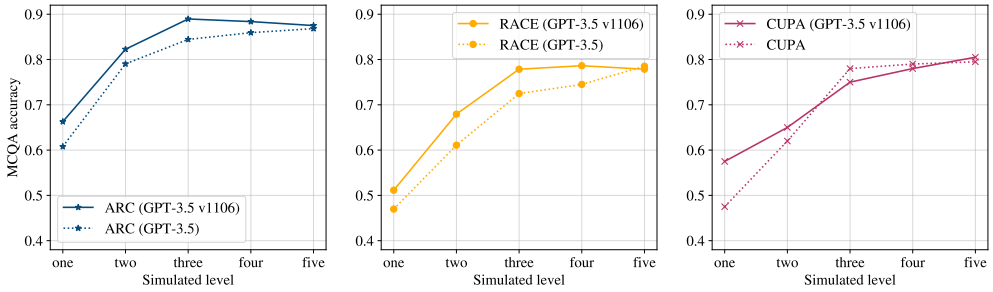

Figure 7: Comparison of *gpt-3.5-turbo-0613* (GPT-3.5) and *gpt-3.5-turbo-1106*, when using the reference prompts.

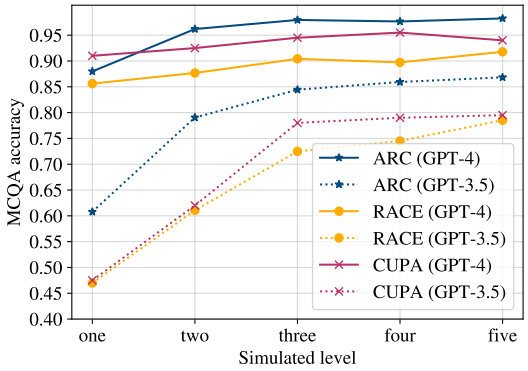

Figure 8: Comparison of the MCQA accuracy of *gpt-3.5-turbo-0613* and *gpt-4-1106-preview* on the three datasets, when prompted with the reference prompts.

scores[13], and TOEFL scores[14]. The prompts are shown in Table 3 in Appendix A.2. Figure 9 shows the experimental results for the three scales on *RACE* and *CUP&A*: in all cases, the model is not capable of accurately simulating students of different proficiency levels, as there is not a correlation between the increase in the student level and the increase in MCQA accuracy. We can however see a difference between CEFR levels, which seem to have an increasing trend up to level B2, and the other scales, which do not show any kind of correlation between the simulated level and the MCQA accuracy. These results suggest that that, although LLMs can simulate students of different levels on abstract scales (such as *one* to *five* in the reference prompt), it is not straightforward to simulate specific proficiency levels. To better analyse this, we also performed additional experiments, adding in the prompt a description of the "meaning" of each proficiency level. For these experiments, as well, we did not find any correlations between the MCQA accuracy and the increase in the role-played level; the complete analysis is available in the Appendix B.2.

4.5.2 Exam grades (marks)

In many educational settings, students are marked with grades on a scale from A (best performing students) to F (lowest performing). We also experiment with (the prompts are shown in Table 4 in Appendix A.3) it and the results for all datasets are shown in Figure 10. The three lines show that, for all datasets, this prompt leads to a model behaviour very close to the desired one, and it is arguably even better than the reference prompts. This is particularly relevant since the LLM does not have view of the whole exam, but answers one question at a time without having information about the others.

5 Related Work

5.1 User Modelling with LLMs

Previous research discussed the possibility of using LLMs instead of (or in addition to) human participants in surveys (Dillion et al., 2023; Argyle et al., 2023; Demszky et al., 2023), and studied whether LLMs can be prompted to show human-like behaviours in a series of task (Aher et al., 2023). However, it is not agreed whether this is actually a good practice. Indeed, some researchers argue that LLMs cannot (and should not) replace human research participants (Harding et al., 2023; Crockett and Messeri, 2023). We mostly agree with the latter, but believe that exam simulations are a different application scenario, as knowledge-based exam questions are built to assess students knowledge in an objective (as much as possible) manner. Still, possible biases of this approach will

[13] https://ielts.org/organisations/ielts-for-organisations/ielts-scoring-in-detail

[14] We use the scores that map to specific IELTS levels in the official documentation: https://www.ets.org/toefl/score-users/ibt/compare-scores.html

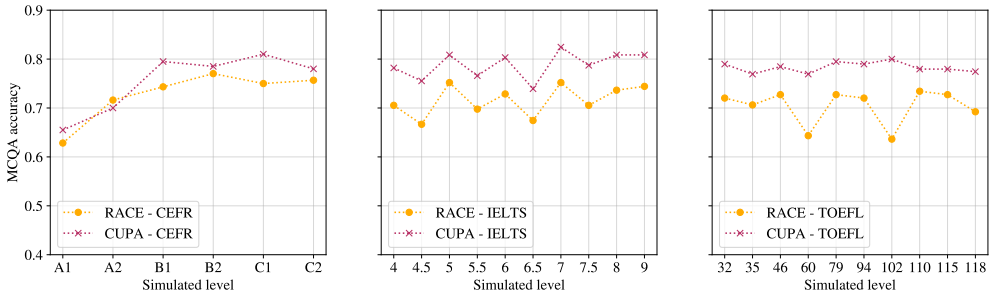

Figure 9: Evaluation of GPT-3.5 when simulating students at specific levels of language proficiency scales.

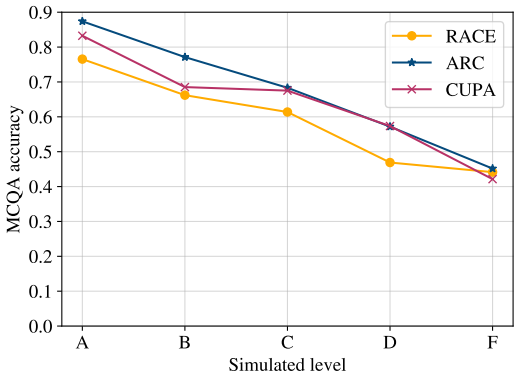

Figure 10: Evaluation of GPT-3.5 while simulating students who got specific grades.

have to be studied before an application in the real world. An approach like the one proposed by Beck et al. (2023), who discusses the possibility of using LLMs as a preliminary step before the human annotations, might be adopted in education, for instance pretesting with human learners only a fraction of the original items.

5.2 LLMs in Education

Previous research discussed profusely the potential of LLMs in education (Jeon and Lee, 2023; Kasneci et al., 2023; Caines et al., 2023). Closer to our work, previous research experimented on Knowledge Tracing with LMs (Liu et al., 2022), but without using them for simulating students. Also related to the current work is the previous research of question difficulty estimation with NLP techniques (AlKhuzaey et al., 2023; Benedetto et al., 2023), especially when performed in an unsupervised manner (Loginova et al., 2021). Indeed, the students simulation we propose in this paper could be used as an alternative to previous approaches for difficulty estimation.

6 Conclusions and future work

In this paper, we have shown that it is possible to prompt GPT-3.5 to *role-play as* (i.e., simulate) students of different levels, and the *reference prompt* we have engineered proved capable of generalising across datasets. However, the actual MCQA accuracy is not easily controllable, and the LLM did not seem capable of representing students at specific levels on standardised language proficiency scales. Crucially, from a practitioner perspective, even though the prompt seems to generalise well to unseen data, it does not seem to generalise to different LLMs: experimenting both with a newer version of GPT-3.5 and GPT-4 we have observed a drift away from the desired behaviour.

Although we found some strong indications that it might be possible to simulate students of different levels with LLMs, there are questions still to be addressed. For a better simulation, one could try to use retrieval augmented generation (RAG) (Lewis et al., 2020) on topic specific documents to better define the level of the role-played student. For a better virtual pretesting, it will be needed to have a larger set of simulated students, possibly increasing the *temperature* of the model (we only use 0) and repeating the simulations several times. Also, it might be helpful to simulate whole exams, instead of one question at a time as we did here.

Future work could also iterate on the *reference prompts*, possibly using automatic prompt optimization (Pryzant et al., 2023), and experiment with other LLMs – preliminary experiments with Llama 2 and Vicuna showed promising results. This is a particularly relevant point since the prompts do not generalise to other LLMs, and specific versions of closed LLMs might be deprecated.

7 Limitations

This work uses LLMs to simulate the responses of students to exam questions and, therefore, any decision taken upon these simulations is at risk of being biased, due to the intrinsic biases in LLMs. This risk is mitigated by the fact that exam questions are built to assess domain knowledge, but are still present. Focusing on the aspects that are specific to the educational domain, it might happen that LLMs reproduce response patterns (and errors) only of a fraction of the population of students, similarly to how using LLMs for surveys oversamples WEIRD[15] participants (Apicella et al., 2020). If this is the case, virtual pretesting done with LLMs would not account for all the other students who make different errors. An example in language learning is the fact that students from different L1s (i.e., first language), tend to make different mistakes. If LLMs reproduce the errors of specific L1s only, this might disadvantage learners with specific backgrounds. This is a common challenge in exam item writing, and even human experts struggle with it. Possible ways to address this are i) to perform pretesting with the desired population of learners and analyse whether their responses are aligned with the ones from the models, and ii) look for biases with the *Marked Personas* approach proposed by Cheng et al. (2023).

An important point that we have raised in this paper is that the results do not seem to generalise across LLMs, as prompts which were very effective on *gpt-3.5-turbo-0613* did not work as well on *gpt-3.5-turbo-1106* and, especially, GPT-4 (*gpt-4-1106-preview*). This is a significant concern from a practitioner's perspective, since any process based on a similar approach might become unusable as soon as there is a new version of the LLM and the older one is deprecated, and suggests that moving towards open LLMs could be a better alternative.

It is worth mentioning that one of the limitations of this approach is the instability of the prompts, and the fact that minor changes to the input prompt might lead to major differences in behaviour. This is a common issue with LLMs, and could be partially mitigated by performing automatic prompt optimization as mentioned in the conclusions.

Lastly, the training dataset of GPT* models is not precisely known, and one might think that this could affect the results shown in this work. Indeed, *ARC* and *RACE* provide some information about question difficulty, and this might be leveraged in some way by the model to adapt its responses to question difficulty. We believe that it is not the case, since the *CUP&A* dataset was released very recently – it is more recent than the training data used in all the models considered in this work – and the finding are consistent across datasets.

[15] Western, Educated, Industrialized, Rich, Democratic.

Acknowledgements

Removed for anonymity.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2023. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, pages 1–53.

Coren Apicella, Ara Norenzayan, and Joseph Henrich. 2020. Beyond weird: A review of the last decade and a look ahead to the global laboratory of the future.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective nlp tasks. *arXiv preprint arXiv:2309.07034*.

Luca Benedetto. 2023. A quantitative study of nlp approaches to question difficulty estimation. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 428–434. Springer Nature Switzerland.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Øistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, et al. 2023. On the application of large language models for language teaching and assessment technology.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Molly Crockett and Lisa Messeri. 2023. Should large language models replace human participants?

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, pages 1–14.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory*, volume 2. Sage.

Jacqueline Harding, William D'Alessandro, NG Laskowski, and Robert Long. 2023. Ai language models cannot replace human research participants. *AI & SOCIETY*, pages 1–3.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, pages 1–20.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In *Asian Conference on Machine Learning*, pages 742–757. PMLR.

Naiming Liu, Zichao Wang, Richard Baraniuk, and Andrew Lan. 2022. Open-ended knowledge tracing for computer science education. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3849–3862.

Ekaterina Loginova, Luca Benedetto, Dries Benoit, and Paolo Cremonesi. 2021. Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 846–855.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark JF Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, et al. 2023. The cambridge multiple-choice questions reading dataset.

Eiji Muraki, Catherine M Hombo, and Yong-Won Lee. 2000. Equating and linking of performance assessments. *Applied Psychological Measurement*, 24(4):325–337.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A List of prompts

A.1 *Analysis of the reference prompt*

Table 2 shows the text of the four prompts that are compared to the reference prompt in Section 4.1 and whose behaviour is shown in Figure 1.

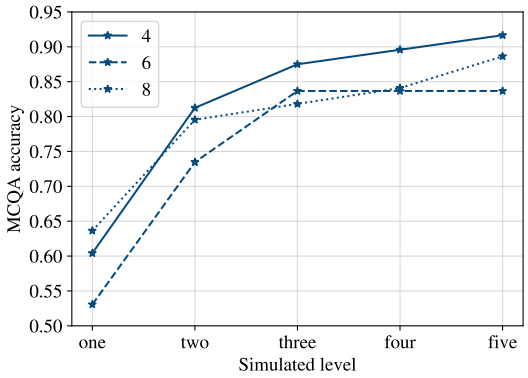

Figure 11: Evaluation of the MCQA accuracy of GPT-3.5 on *ARC* when simulating students of different levels, separately on questions of different *grades*.

A.2 *Language proficiency scales*

Table 3 shows the prompts that are used for the experiments on language proficiency scales analysed in Figure 9 in Section 4.5.

A.3 *Exam grades (marks)*

Table 4 shows the prompts that are used for the experiments on exam marks analysed in Figure 10 in Section 4.5.2.

B Additional analyses

B.1 *ARC*: MCQA accuracy per *grade*

Figure 11 complements Figure 5 (in Section 4.2.1) by showing, separately for the even question *grades*, the MCQA accuracy of GPT-3.5 when simulating students of different levels. The figure shows that the behaviour of the reference prompt is similar at what was observed in the other analyses: there is a trend of increasing MCQA accuracy for increasing simulated levels but, in this case, it is not true that the most difficult questions (*grade* 8), lead to the lowest accuracy.

B.2 *RACE* and *CUP&A*: additional analyses on language proficiency scales

Figure 12 shows the additional analyses on language proficiency scales. Specifically, we perform additional modifications to the reference prompt, to explore its understanding of the language proficiency scales under consideration. The list of prompts is shown in Table 5. The Figure shows that, for all language proficiency scales and prompts considered, there is not a clear correlation between the increase in the simulated student level and the

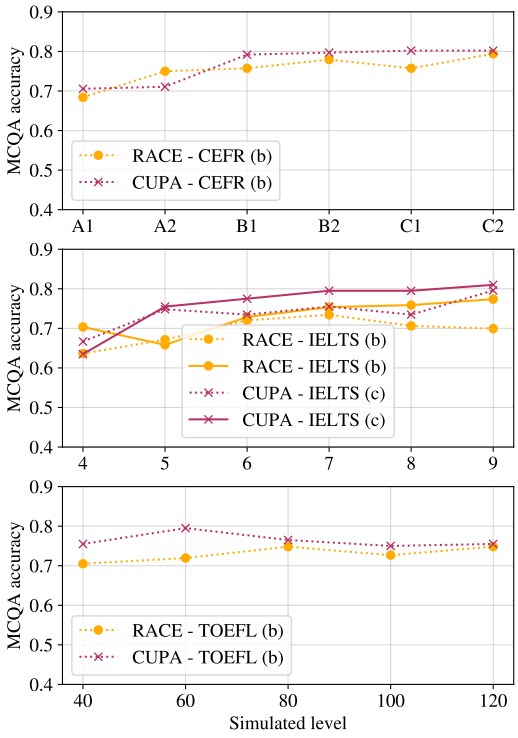

Figure 12: Evaluation of GPT-3.5 when representing students at specific levels of language proficiency scales, *RACE* and *CUP&A* datasets.

MCQA accuracy, not even when we explicitly describe the capabilities that a learner of a specific level should have (prompt CEFR (b)).

B.3 Abstract scale: *beginner*, *intermediate*, *advanced*

In addition to experimenting with standardised educational scale, we also evaluate the behaviour of GPT-3.5 on another *abstract* scale. Specifically, we consider the student levels *beginner*, *intermediate*, *advanced*, and do not provide any additional information to the model. The results are shown in Figure 13. In this setting, the model is asked to represent only three levels, which is arguably an easier task. Still, for all datasets, we can observe the desired monotonic trend of increasing MCQA accuracy for increasing simulated levels. This suggests that indeed GPT-3.5 is capable of simulating students at different levels, and this seems easier to do with abstract scales.

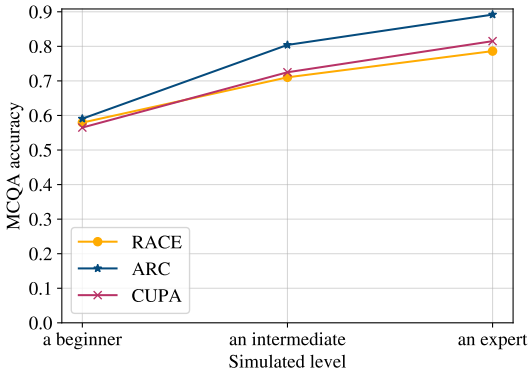

Figure 13: Evaluation of GPT-3.5 when simulating students of *beginner*, *intermediate*, and *advanced* levels on the three datasets.

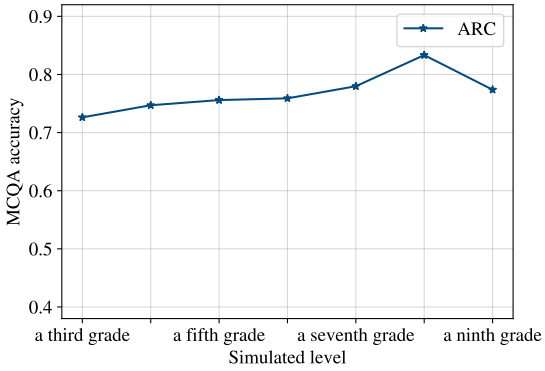

Figure 14: Evaluation of GPT-3.5 when simulating students of different school grades.

B.4 School grades

We perform this experiment on *ARC* only, since each question is associated with a *grade*, and we ask the LLM to simulate students from third grade to ninth grade, which are the grades in *ARC*. The results, shown in Figure 14, suggest that although there seems to be a correlation between the increase in school grade and MCQA accuracy, this is not consistent (there is a drop for ninth graders) and the difference in accuracy between highest and lowest level students is fairly small and probably not representative of real students.

Table 2: List of prompts showed in Figure 1 in Section 4.1, where they are compared with the reference prompt. For all prompts, the student levels we consider are `['one', 'two', 'three', 'four', 'five']`.

ID	Prompt
28	SYSTEM: You will be shown multiple choice questions from a science exam. The questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). Similarly, the students can be identified with skill levels from one (low level student) to five (very skilled student). The level of students is defined such that a student of a certain level can answer most of the questions of lower levels, and almost none of the question of higher levels. You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of level {X} would pick. Provide only a JSON file with the following structure: {“level”: “difficulty level of the question”, “index”: “integer index of the answer chosen by a student of level {X}”, “text”: “text of the answer chosen by the student”} USER: Question: “{question}” Options: “{answer options}”
31	SYSTEM: You will be shown multiple choice questions from a science exam. The questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of level {X} would pick. Provide only a JSON file with the following structure: {“level”: “difficulty level of the question”, “index”: “integer index of the answer chosen by a student of level {X}”, “text”: “text of the chosen answer”} USER: Question: “{question}” Options: “{answer options}”
32	SYSTEM: You will be shown multiple choice questions from a science exam. The questions in the exam have difficulty levels on a scale from level one (very easy) to level five (very difficult). Similarly, each student can be given a skill level: level one represents the least skilled students, who answer most questions wrongly, and level five represents the most skilled students, who can correctly answer even the most difficult items. You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of level {X} would pick. Provide only a JSON file with the following structure: {“level”: “difficulty level of the question”, “index”: “integer index of the answer chosen by a student of level {X}”, “text”: “text of the chosen answer”} USER: Question: “{question}” Options: “{answer options}”
35	SYSTEM: You will be shown multiple choice questions from a science exam. The questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, motivating your choice, and select the answer choice that a student of level {X} would pick. Provide only a JSON file with the following structure: {“level”: “difficulty level of the question”, "motivation": "reason why you assigned that difficulty level", “index”: “integer index of the answer chosen by a student of level {X}”, “text”: “text of the chosen answer”} USER: Question: “{question}” Options: “{answer options}”

Table 3: List of prompts used for the experiments on language proficiency scales analysed in Figure 9 in Section 4.5. These prompts are evaluated on *RACE* and *CUP&A* only. In **bold** the parts that are different from the *reference prompts*.

ID	Prompt	Student levels
44	SYSTEM: You will be shown a multiple choice question from an English reading comprehension exam, and the questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of **CEFR** level {X} would pick. Provide only a JSON file with the following structure: {"question level": "difficulty level of the question", "answer explanation": "the list of steps that the students of **CEFR** level {X} would follow to select the answer, including the misconceptions that might cause them to make mistakes", "index": "integer index of the answer chosen by a student of **CEFR** level {X}"} USER: Reading passage: "{context}" Question: "{question}" Options: "{answer options}"	[A1, A2, B1, B2, C1, C2]
45	SYSTEM: You will be shown a multiple choice question from an English reading comprehension exam, and the questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of **IELTS** level {X} would pick. Provide only a JSON file with the following structure: {"question level": "difficulty level of the question", "answer explanation": "the list of steps that the students of **IELTS** level {X} would follow to select the answer, including the misconceptions that might cause them to make mistakes", "index": "integer index of the answer chosen by a student of **IELTS** level {X}"} USER: Reading passage: "{context}" Question: "{question}" Options: "{answer options}"	[4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 9]
46	SYSTEM: You will be shown a multiple choice question from an English reading comprehension exam, and the questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of **TOEFL** level {X} would pick. Provide only a JSON file with the following structure: {"question level": "difficulty level of the question", "answer explanation": "the list of steps that the students of **TOEFL** level {X} would follow to select the answer, including the misconceptions that might cause them to make mistakes", "index": "integer index of the answer chosen by a student of **TOEFL** level {X}"} USER: Reading passage: "{context}" Question: "{question}" Options: "{answer options}"	[32, 35, 46, 60, 79, 94, 102, 110, 115, 118]

Table 4: Prompts used for the experiments on exam marks analysed in Figure 10 in Section 4.5.2. The first prompt is used on *ARC*, the second on *RACE* and *CUP&A*. In **bold** the parts that are different from the *reference prompts*.

ID	Prompt	Student levels
55	SYSTEM: You will be shown a multiple choice question from a science exam, and the questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a **grade {X} student** would pick. Provide only a JSON file with the following structure: {"question level": "difficulty level of the question", "answer explanation": "the list of steps that a **grade {X} student** would follow to select the answer, including the misconceptions that might cause them to make mistakes", "index": "integer index of the answer chosen by a **grade {X} student**"} USER: Question: "{question}" Options: "{answer options}"	[A, B, C, D, F]
57	SYSTEM: You will be shown a multiple choice question from an English reading comprehension exam, and the questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a **grade {X} student** would pick. Provide only a JSON file with the following structure: {"question level": "difficulty level of the question", "answer explanation": "the list of steps that a **grade {X} student** would follow to select the answer, including the misconceptions that might cause them to make mistakes", "index": "integer index of the answer chosen by a **grade {X} student**"} USER: Reading passage: "{context}" Question: "{question}" Options: "{answer options}"	[A, B, C, D, F]

Table 5: List of prompts used for the additional experiments on language proficiency scales. The results are shown in Figure 12. In the table we show the final part of the prompt, (starting with "Provide only a JSON file with...") and the USER message only for the first prompt as they are the same for all prompts. The CEFR descriptions mentioned in prompt *CEFR (b)* are taken from the "Common Reference levels: Global scale" provided by the council of Europe. The short descriptions for the IELTS levels used in prompts *IELTS (b)* and *IELTS(c)* are taken from the official IELTS website.

ID	Prompt	Student levels
CEFR (b)	SYSTEM: You will be shown multiple choice questions from a science exam. The questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of CEFR level {X} would pick. A student of CEFR level {X} can {CEFR level description}. Provide only a JSON file with the following structure: {"level": "difficulty level of the question", "index": "integer index of the answer chosen by a student of CEFR level {X}", "text": "text of the chosen answer"} USER: Reading passage: "{context}" Question: "{question}" Options: "{answer options}"	[A1, A2, B1, B2, C1, C2]
IELTS (b)	SYSTEM: You will be shown a multiple choice question from an English reading comprehension exam, and the questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of IELTS level {X} would pick. The meaning of the IELTS levels is as follows: - IELTS level 9 indicates an Expert test taker; - IELTS level 8 indicates a Very good test taker; - IELTS level 7 indicates a Good test taker; - IELTS level 6 indicates a Competent test taker; - IELTS level 5 indicates a Modest test taker; - IELTS level 4 indicates a Limited test taker; Provide only a JSON file ...	[4, 5, 6, 7, 8, 9]
IELTS (c)	SYSTEM: You will be shown multiple choice questions from a science exam. The questions in the exam have difficulty levels on a scale from level one (very easy) to level five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of IELTS level {X} ({IELTS level short description}) would pick. Provide only a JSON file ...	[4, 5, 6, 7, 8, 9]
TOEFL (b)	SYSTEM: You will be shown multiple choice questions from a science exam. The questions in the exam have difficulty levels on a scale from one (very easy) to five (very difficult). You must assign a difficulty level to the given multiple choice question, and select the answer choice that a student of TOEFL level {X} would pick. Provide only a JSON file ...	[40, 60, 80, 100, 120]