NLP Methods May Actually Be Better Than Professors at Estimating Question Difficulty

Leonidas Zotos[1,3,*], Ivo Pascal de Jong[2,*], Matias Valdenegro-Toro[2], Andreea Ioana Sburlea[2], Malvina Nissim[1] and Hedderik van Rijn[3]

[1]Center for Language and Cognition, University of Groningen
[2]Bernoulli Institute, University of Groningen
[3]Department of Experimental Psychology, University of Groningen

Abstract
Estimating the difficulty of exam questions is essential for developing good exams, but professors are not always good at this task. We compare various Large Language Model-based methods with three professors in their ability to estimate what percentage of students will give correct answers on True/False exam questions in the areas of Neural Networks and Machine Learning. Our results show that the professors have limited ability to distinguish between easy and difficult questions and that they are outperformed by directly asking `Gemini 2.5` to solve this task. Yet, we obtained even better results using uncertainties of the LLMs solving the questions in a supervised learning setting, using only 42 training samples. We conclude that supervised learning using LLM uncertainty can help professors better estimate the difficulty of exam questions, improving the quality of assessment.

Keywords
item difficulty estimation, uncertainty estimation, educational data, large language models, assessment

1. Introduction

Good exam design is time-consuming and difficult. One of the challenges is to ensure consistent difficulty over multiple years, as exam scores should be comparable between cohorts. As previous exams might circulate among students, instructors are required to design exams anew, selecting questions that are neither too difficult, nor too easy [1]. One solution is to randomly select a sufficiently large sample of questions from an *item pool* [2]. However, the number of questions can often not be sufficiently large to be confident that the difficulty will remain constant over years. This requires instructors to estimate the difficulty of the questions to ensure consistency, a process that is often an implicit aspect of exam design.

In this paper we assess whether Artificial Intelligence (AI), and in particular Natural Language Processing (NLP), can be used to assist instructors in this process. AI is being viewed as a valuable avenue for decreasing workload and increasing the capacity of educational staff in a variety of applications, ranging from tutor chat-bots to systems that can grade exams [3]. Even though using AI for difficulty estimation has been explored [4], success has been modest [5], with NLP systems often performing marginally better than average-based baselines [6]. This task is also challenging for teachers, as shown by van de Watering and van der Rijt [7]. They found that teachers could correctly estimate the difficulty levels for only a small proportion of the questions.

The modest success of question difficulty estimation using NLP methods and the known limitations of teachers to estimate question difficulty motivates this study. It is clear that both teachers and NLP-based methods have a limited ability to estimate exam item difficulty, but it is not known how they compare. This comparison is critical to determining whether automated question difficulty estimation is ready for educational practice. In our work, we compare NLP-based approaches for automated question difficulty estimation with expert-human estimation of difficulty. We demonstrate that state-of-the-art

2nd International Workshop on AI in Society, Education and Educational Research (AISEER), October 25–30, 2025, Bologna, Italy
*Corresponding authors; These authors contributed equally to this work.
l.zotos@rug.nl (L. Zotos); ivo.de.jong@rug.nl (I. P. d. Jong)

NLP methods are better at question difficulty estimations than university professors, and highlight the potential of integrating NLP-based methods in the workflow of exam design.

2. Related Work

The task of question difficulty estimation using NLP methods is not new [4]. Already in the 1990s, traditional AI methods, were employed for question difficulty estimation [8, 9]. More recent approaches are typically based on the transformer architecture. An example of this is the recent "Building Educational Applications" shared task on "Automated Prediction of Item Difficulty and Item Response Time" [6], wherein a variety of approaches were explored ranging from changing the transformer architecture to data augmentation techniques. The best performing team (EduTec) used a combination of model optimisation techniques including scalar mixing [10], rational activation [11] and multi-task learning to predict the proportion of students answering each question correctly [12].

Our Contribution

The goal of this study is to establish whether modern NLP-based methods can be applied for question difficulty estimation in university education. This is operationalized by comparing whether NLP-based approaches perform similar or better than the lecturers who would normally construct the exams. To the best of our knowledge, this is the first study of this kind. This comparison is conducted using two university-level exams, the moderate-size question set being representative of the data that would typically be available in real-world scenarios. The code implementation of this project is publicly available.[1]

3. Methods

To compare the performance of professors and LLM-based solutions in question difficulty estimation, we collected a dataset of exam questions used in university education. The proportion of students answering a question correctly (known as the p^+-value) is considered the ground-truth difficulty. The professors and LLM-based methods estimate this ground truth based on the exam question text. We chose to use the p^+-value over IRT metrics [13], because it is more intuitive for a professor to interpret and estimate.

3.1. Exam data

We included data from two courses in the area of Artificial Intelligence that are taught at University of Groningen. Specifically, we used Neural Networks, which is taught in the Artificial Intelligence BSc program, and Advanced Machine Learning, which is taught in the Artificial Intelligence MSc program. Both courses and exams follow a similar setup. The course material consists of custom lecture notes, and the exams are made of twenty-two questions selected from a private item pool of exam questions. Each question in the exam is a True/False question, and students have two hours to complete the exam. Examples of exam questions are shown in Table 1.

We collected three archived exams for each course, covering the years 22/23 (111 students), 23/24 (114 students), and 24/25 (20 students) for Neural Networks and the years 21/22 (103 students), 22/23 (119 students) and 23/24 (71 students) for Advanced Machine Learning. We collected all questions from the exams and pooled them together. For questions that were repeated across years, the p^+-value was based on all students that received this question. This was the case for five questions in total for each of the two courses. Moreover, the examiner of the courses considered four questions from Advanced Machine Learning and one question from Neural Networks as ambiguous and marked those as correct for both True and False. Those ambiguous questions were removed from this study. Additionally, there

[1] https://github.com/LeonidasZotos/nlp_vs_professors_difficulty_estimation

Table 1
Two example exam questions which were given to the students as practice material.

Exam Question	Answer
(Machine Learning basics) Given a training data set $(\mathbf{u}_i, \mathbf{y}_i)_{i=1,\ldots,N}$, where $\mathbf{u}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in \mathbb{R}^m$, then for any model $f : \mathbb{R}^n \to \mathbb{R}^m$ and any loss function L, the empirical risk $\mathcal{R}^{emp}(f)$ is less or equal to the risk $\mathcal{R}(f)$.	False
(Elementary math) Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be differentiable functions with gradients $\nabla f, \nabla g$. Then $\nabla(f + g) = \nabla f + \nabla g$.	True

was one question which included an image. This question was also removed as we consider this to be out-of-scope for the present study. This resulted in 59 questions from Neural Networks and 53 questions from Advanced Machine Learning.

We use this new dataset instead of using existing datasets for three reasons. First and foremost, in contrast to other datasets, we have professors that are experts in the field available that can provide p^+-value estimates to represent the manual question difficulty estimation in a way that is ecologically valid. Secondly, the questions in our new dataset are not publicly available guaranteeing that they are unseen for all LLMs. Finally, by analyzing how difficulty estimation methods perform on questions involving abstract mathematical reasoning and comprehension, we examine whether their previous success in assessing the difficulty of clinical decision-making and language comprehension exams [6, 14] extends to this domain.

3.2. Professors' estimations

Three professors of the University of Groningen were asked to estimate for each question the percentage of students that would answer correctly. All three professors have expertise in Machine Learning and Neural Networks and would be qualified to teach these courses. However, none of them have taught these specific courses, and have never been a student in these courses. This ensures they are fairly knowledgeable about the population of students and the topic, but have not seen students' performance on these questions. As an example and to provide some calibration, the professors were given one exam question with the true percentage of students that answered it correctly. The professors were also given the correct True/False answers. This was to help them focus on the task — predicting the difficulty of the question rather than solving it. The exact annotation instructions are presented in Table 2.

Table 2
Instructions given to the professors. Equivalent instructions were given for the Neural Networks exam.

Annotation Instruction: Below are exam questions from the Advanced Machine Learning Course, taught in the University of Groningen. For each question, the correct answer is highlighted in **green**. Estimate, from the examiner's perspective, what percentage of students will answer each question correctly. Feel free to re-visit and adjust previous estimates. An example is presented below, where the percentage of students who selected the correct answer is provided. At the end, provide an estimate of the time spent on this item difficulty estimation task. **Purpose of the Study:** This study aims to compare the performance of expert educators and state-of-the-art LLM-based methods in estimating the difficulty of True/False questions.

Each professor made their estimates independently and at a moment that fits their schedule. On average, estimating the difficulty of the total of 112 questions took each professor 2 hours and 15 minutes. One professor (professor 3) declined to give estimates for sixteen questions for Neural Networks and six questions for Advanced Machine Learning, stating that they *miss the specific knowledge of some concepts to provide a confident estimate.* These questions were not considered in the evaluation for this professor, but were maintained for the rest of the analysis.

3.3. NLP Approaches

We focus on two types of NLP-based methods for item difficulty estimation. We investigate methods based on prompting where LLMs directly estimate the question difficulty and methods based on the uncertainty of an LLM attempting to solve the question. The mathematical notation in all questions is encoded using LaTeX, which LLMs can process well [15, 16].

Using Direct Estimation As a simple comparison between LLMs and professors, we tested two setups in which a powerful LLM is prompted to directly estimate the p^+-value of a question. We provided the LLMs with the same instruction (and example question) that we also gave to the professors. Additionally, we use Chain of Thought (by instructing the LLM to "Think step by step") to allow the model to "reason" before giving an estimate [17]. To get the most competitive results we use `gemini-2.5-pro-preview-03-25` (`Gemini 2.5`) and `Gemini-2.0-flash` (`Gemini 2.0`), two of the best-performing current LLMs, as measured by the community-driven Chatbot Arena [18]. At the time of writing, they rank 1st and 8th respectively.

The LLMs are prompted using two different setups. In the single question setup, the LLM is tasked with predicting the p^+-value of each question individually without being able to see the other questions. In contrast, in the all-questions setup the LLM is given the complete question set and is tasked with estimating the p^+-values of all items in one go. We consider both of these setups to be promising, each with its own trade-offs. On one hand, we observe that prompting the model to generate the difficulty of a single question item encourages it to generate longer reasoning streams, which could lead to more accurate predictions. At the same time, predicting the difficulty of all items concurrently can also be beneficial, potentially steering the prediction of each item to be informed by the entire set. The all-questions setup closely resembles the setup with the professors, as they can also see all questions to gauge the overall difficulty of the question set.

Using LLM Uncertainty As a task-specific question difficulty estimation method we implement the approach by Zotos et al. [19], which is a good representation of the current state-of-the-art for this task.

In this approach, a set of nine LLMs are prompted to solve each question without the answer, where the uncertainty of the LLM can be used as a feature for a supervised learning model to predict the p^+-value. By using a mixture of stronger and weaker LLMs we can get a good spread of LLM uncertainties as features. We use the same LLMs as in the original work.

Two measures of uncertainty are used to indicate the difficulty of the question according to the LLM. One is the probability of the first generated token (probability of "A" or "B"). The other one is Choice-Order Sensitivity [20], which measures whether the LLM gives the same prediction when the order of the answer choices is shuffled. This is operationalised by performing inference with different permutations of choice-order (for True/False questions, only two permutations are possible) and calculating the proportion of times the correct answer is selected. Both measures have been found to correlate with the probability that a prediction from an LLM is correct [21, 19] as well as the p^+-values of exam questions [22].

As supervised learning models we use three different regressors: Random Forest, Support Vector Machine and Linear Regression. Each model is trained for each course using an 80:20 train-test split. The regression models for Neural Networks are therefore trained with 47 samples, and the models for Advanced Machine Learning with 42 samples. Using Grid Search with 5-fold validation we determine the best hyperparameters for each regression model.

We compare this with two alternative supervised learning setups. In one we train a dummy model with no features, always predicting the mean p^+-value. In the other, we consider what may be learned with simple features from the text. For this we use TF-IDF features for a supervised learning model. For completeness, we also consider the concatenation of TF-IDF features and LLM uncertainties as features.

Arguably, comparing the professors to this supervised learning approach is unfair. The professors are only given one "labeled" example to calibrate their predictions, while the regressor needs more than one example to undergo training. The supervised-learning approach therefore can directly learn the

Table 3
Root Mean Squared error (RMSE), Mean Error (ME) and Spearman's ρ as measured for each Item Difficulty Estimation Approach. Overall best performance is indicated with **boldface**, while *italics* indicates best in category.

Method	Features	NN			AML		
		RMSE ↓	ME	ρ ↑	RMSE ↓	ME	ρ ↑
Professors							
Professor 1		0.336	0.048	-0.014	0.357	-0.087	0.062
Professor 2		0.245	-0.038	-0.011	*0.185*	0.070	*0.241*
Professor 3		*0.184*	0.000	*0.211*	0.193	-0.003	0.224
Average of Professors' Estimates		0.234	0.001	-0.020	0.205	-0.010	0.173
LLM Direct Estimation							
`Gemini 2.0`	Single question	0.242	-0.071	0.118	0.199	-0.061	0.188
	All questions	0.218	0.016	0.062	0.307	-0.177	0.052
`Gemini 2.5`	Single question	*0.192*	0.035	*0.283*	*0.165*	0.044	*0.345*
	All questions	0.199	0.034	0.175	0.243	0.051	0.079
Supervised Learning							
Dummy Model	None	0.174	-	-	0.126	-	-
Linear Regression	LLM Uncertainties	0.165	0.024	0.678	0.119	0.004	0.418
	TF-IDF	0.174	0.016	0.098	0.126	-0.000	0.251
	TF-IDF & Uncertainties	0.165	0.024	0.741	0.119	0.003	0.464
Random Forest	LLM Uncertainties	0.170	0.032	0.273	0.120	-0.009	0.282
	TF-IDF	0.180	0.004	0.035	0.115	0.003	0.469
	TF-IDF & Uncertainties	0.179	0.011	0.049	0.109	0.004	0.464
Support Vector Machine	LLM Uncertainties	0.148	0.011	0.811	**0.107**	0.009	**0.582**
	TF-IDF	0.174	0.018	-0.070	0.117	0.008	0.314
	TF-IDF & Uncertainties	**0.147**	0.001	**0.853**	0.109	0.013	0.491

distribution of p^+-values, which the professors do not have access to. At the same time, this setup is realistic for an advanced NLP setup that may be used in practice. Universities often have archived data of previous years' exams, but reviewing them is time-consuming. Using this supervised learning setup, we can capitalize on this existing data effectively.

4. Results

To evaluate how the professors compare to the NLP-based methods on the question difficulty estimation, we use the root mean squared error (RMSE) between the estimated p^+-values and the ground truth values as a standard metric for error. We also measure the rank correlation between the estimates and the ground truth using Spearman's ρ. This rank correlation assessment allows us to detect whether an approach which has consistently biased estimates still maintains a strong monotonic relationship with the true p^+-values and is thus able to distinguish easy from difficult questions. The Mean Error (ME) metric directly evaluates any consistent bias, by measuring whether the difficulty estimates are on average too high or too low. The results of all experiments are presented in Table 3.

Professor Performance Overall, and in line with the study by van de Watering and van der Rijt [7], professors seem to have limited ability to estimate question difficulty. We see that for the Neural Networks (NN) exam, two of the professors estimated p^+-values that do not correlate with the performance of students. Only professor 3 has a positive rank correlation with ρ=0.211. This may be partly because professor 3 did not give an answer to sixteen questions for Neural Networks, presumably the ones they felt uncertain about. For the MSc level Advanced Machine Learning (AML) course professor 2 and professor 3 achieved better performances. Their estimated p^+-values consistently show a weak rank correlation with the ground truth.

The Mean Errors are sometimes positive and sometimes negative, depending on the professor and the exam. This suggests that there is no clear pattern of professors consistently over/underestimating question difficulty. The high RMSE does show that overall the professors perform poorly at directly estimating p^+-values. Furthermore, for each question we also average the three professor estimates as if they are voting. This did not lead to any improvements.

Direct Prompting Performance When assessing the two methods of direct prompting, we find that prompting the model with one question at a time generally leads to lower RMSE and higher rank correlations, with the exception of `Gemini 2.0` in the Neural Networks set. Additionally, we find that `Gemini 2.5` is consistently more accurate than `Gemini 2.0` which corresponds well with their performance in other tasks [18]. When comparing the direct prompting of LLMs to the professors we find that the LLMs tend to perform better. The best LLM method (`Gemini 2.5`, single question) has better rank correlation than all professors on both exams. For Advanced Machine Learning the best rank correlation from a professor was ρ=0.241, while that of the LLM was ρ=0.345.

Supervised Learning Performance The supervised learning methods achieve lower RMSE than the professors and the direct LLM predictions, because only the supervised learning methods are able to learn the distribution of p^+-values from data. The SVM performs best likely due to the small dataset and non-linear relationships. Using only TF-IDF features was not sufficient to estimate p^+-values for the Neural Networks set, but was already better than the professors and often better than the LLMs for the Advanced Machine Learning set. The LLM uncertainties as features are substantially more predictive, resulting in lower RMSE and higher rank correlation ρ. The SVM with TF-IDF Scores and LLM Uncertainties performed the best, with a rank correlation of ρ=0.853 for Neural Networks. For Advanced Machine Learning, the SVM trained only on LLM Uncertainties performed best, with a rank correlation of ρ=0.582. This is much better than either direct estimation from the LLM, or estimation from the professors.

4.1. Inter-Annotator Agreement

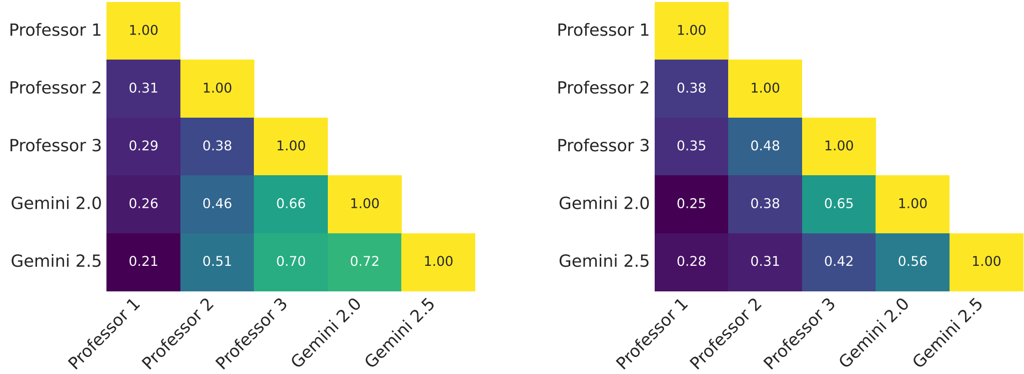
Figure 1 shows the inter-annotator agreement (including the direct assessment of the Gemini LLMs), represented using the Spearman correlation coefficient. Overall, this analysis shows that, while the task is difficult (as was shown earlier), there are moderate correlations between the professors, indicating that they might be over/under-estimating the difficulty of the same questions. This suggests that for Advanced Machine Learning professors have a consistent notion of what should be difficult and what should be easy and are making informed estimates.

Additionally, we observe a high correlation between professor 3 and the Gemini Models in the Neural networks dataset, in line with their relatively good performance on the set. Lastly, we also find a moderate to high correlation between the assessments of the two Gemini LLMs, suggesting that LLMs of the same family behave consistently on this task.

4.2. Per Question Analysis

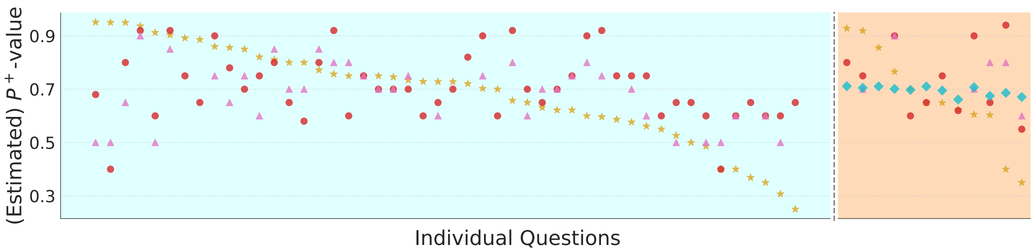
Figure 2 presents, per question, the p^+-value, along with the estimates of the best performing systems per category. For each dataset, we separate the questions based on the train and test splits used for the best-performing Supervised Learning approach (this separation has no impact on the teachers and prompted LLMs). Here, we directly observe that there is a good range and distribution of difficulties, with a balance of easy and difficult questions. We also see that a few questions in each set were answered correctly by less than 40% of the student population, suggesting that these questions might be misleading or trick-questions.

Looking at the general picture, and in line with the results presented in Table 3, the best professors' predictions and `Gemini 2.5`'s predictions do not show a strong correlation with the true p^+-values. Additionally, their estimates show high variability, but are seldom below 50%, suggesting that they do not recognize trick-questions that might lead students to perform worse than random guessing. In

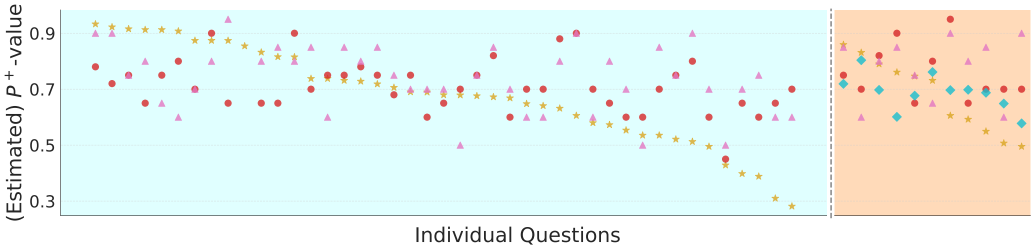


(a) Neural Networks (b) Advanced Machine Learning

Figure 1: Inter-Annotator Agreement, calculated using Spearman's ρ for the two datasets.



(a) Neural Networks



(b) Advanced Machine Learning

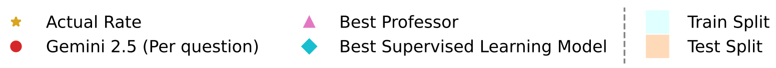


Figure 2: Estimated p^+-value per question item for the best-performing professor, `Gemini 2.5` and trained support vector machine using LLM Uncertainties. The Train/Test split is only relevant for the Supervised Learning Model, for which we only report its performance on the unseen question items.

contrast, the estimated p^+-values of the best trained Supervised Learning Model show low variability and remain near the average p^+-value observed in the training sets. We do see that the estimated p^+-values correlate with the true p^+-values, but that the very easy/difficult questions are estimated close to the average. This explains the good rank correlation, but still high RMSE that we observed in Table 3.

5. Discussion

We have shown that `Gemini 2.5` is better at question difficulty estimation for Neural Networks and Advanced Machine Learning exams than three university professors. We also find that `Gemini 2.5` consistently outperforms `Gemini 2.0` on this task, which suggests that future LLM releases may lead to further improvement.

Additionally, we find that with as little as 42 training samples the supervised learning method of Zotos et al. [22] based on the uncertainty of LLMs solving the problem substantially outperforms both professors and a standard LLM approach. This finding is significant, as it demonstrates that implementing this system for individual courses is feasible, with only a couple of exams from previous years being required to train a good regression model.

Lastly, our findings are on questions that require parsing mathematical notation and that require mathematical reasoning. This extends previous successes of NLP-based question difficulty estimation on biopsychology [22], clinical decision making [6] and language comprehension [14] exams to more mathematical fields. While the current results are specific to Machine Learning, they suggest that NLP methods are also promising to other mathematical topics such as physics, computer science and astronomy.

Overall we have demonstrated that state-of-the-art NLP methods are – relative to professors – very good at question difficulty estimation, and can support them in ranking question difficulty. Of course our findings are focused on question difficulty estimation, and we still need professors for the many other aspects of exam design and education!

Limitations The primary limitation of this study is that we relied on professors that did not teach this specific course. In reality, professors have additional information which can help with the task of question difficulty estimation, such as the performance of a cohort during the semester. At the same time, the professors in our study already have significantly more background information than the better-performing `Gemini 2.5`, as they are familiar with the rest of the curriculum and know how these students perform in other courses. This limitation may cast doubt whether question difficulty estimation from `Gemini 2.5` is better than a professor that has been teaching a specific course. However, it remains clear that the supervised learning method is superior given the large differences.

We also observed that professors mostly make p^+-value predictions in increments of 5% (e.g., 65% or 70%, but not 68%). This results in items being tied in terms of predicted p^+-value. While these ties do not impact the calculation of the Root Mean Squared Error, they might negatively affect the Spearman Rank Correlation Coefficient, as granular judgments that would create a clear ranking of the question items is not available. However, we observe that `Gemini 2.5`'s direct estimations also frequently occur in 5% increments (with the model often predicting 60% or 75%, as shown in Figure 2), yet a consistently higher correlation is observed compared to the professors' annotations.

Broader Impact Statement While the results of the current study are promising, implementing such a system is not trivial: The best performing system we tested relies on the existence of some training data, as well as the availability of sufficient computational resources to compute the uncertainty metrics of the LLMs. At the same time, instructing a state-of-the-art proprietary LLM to estimate question difficulty can lead to good performance on the task, a solution that is trivial to use. As a final consideration, we believe that any system of this type should be used in a human-in-the-loop fashion to address cases where the NLP methods unavoidably lack context such as, for example, when a question is assessed as easy even though the material was not covered in class.

Acknowledgments

We would like to thank Professor Herbert Jaeger for providing the exam material for this study. We would also like to thank the three professors for volunteering their time and expertise to provide estimates of question difficulty. This study was positively reviewed by the Faculty of Science and Engineering Ethics Committee under reference FSE.EC25005.

Declaration on Generative AI

For the preparation of this work, Gemini 2.5 was used for: Grammar check.

References

[1] L. F. Bachman, Fundamental considerations in language testing, Oxford university press, 1990.

[2] W. D. Way, Protecting the integrity of computerized testing item pools, Educational Measurement: Issues and Practice 17 (1998) 17–27.

[3] W. Holmes, I. Tuomi, State of the art and practice in AI in education, European journal of education 57 (2022) 542–570.

[4] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, A survey on recent approaches to question difficulty estimation from text, ACM Computing Surveys 55 (2023) 1–37.

[5] S. AlKhuzaey, F. Grasso, T. R. Payne, V. Tamma, Text-based question difficulty prediction: A systematic review of automatic approaches, International Journal of Artificial Intelligence in Education 34 (2024) 862–914. doi:10.1007/s40593-023-00362-1.

[6] V. Yaneva, K. North, P. Baldwin, L. A. Ha, S. Rezayi, Y. Zhou, S. R. Choudhury, P. Harik, B. Clauser, Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions, in: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), 2024, pp. 470–482.

[7] G. van de Watering, J. van der Rijt, Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items, Educational Research Review 1 (2006) 133–147. URL: https://www.learntechlib.org/p/197391.

[8] K. Perkins, L. Gupta, R. Tammana, Predicting item difficulty in a reading comprehension test with an artificial neural network, Language Testing 12 (1995) 34–53. doi:10.1177/026553229501200103.

[9] R. F. Boldt, R. Freedle, Using a neural net to predict item difficulty, ETS Research Report Series (1996) i–19. doi:https://doi.org/10.1002/j.2333-8504.1996.tb01709.x.

[10] S. Gombert, D. D. Mitri, O. Karademir, M. Kubsch, H. Kolbe, S. Tautz, A. Grimm, I. Bohm, K. Neumann, H. Drachsler, Coding energy knowledge in constructed responses with explainable NLP models, Journal of Computer Assisted Learning 39 (2023) 767–786.

[11] A. Molina, P. Schramowski, K. Kersting, Padé activation units: End-to-end learning of flexible activation functions in deep networks, in: 8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=BJlBSkHtDS.

[12] S. Gombert, L. Menzel, D. D. Mitri, H. Drachsler, Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads, in: E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, Z. Yuan (Eds.), Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 483–492. URL: https://aclanthology.org/2024.bea-1.40.

[13] R. K. Hambleton, H. Swaminathan, H. J. Rogers, Fundamentals of item response theory, volume 2, Sage, 1991.

[14] A. Mullooly, Ø. Andersen, L. Benedetto, P. Buttery, A. Caines, M. J. F. Gales, Y. Karatay, K. Knill, A. Liusie, V. Raina, S. Taslimipoor, The Cambridge Multiple-Choice Questions Reading Dataset, Cambridge University Press and Assessment, 2023. doi:10.17863/CAM.102185.

[15] S. Frieder, L. Pinchetti, C. Chevalier, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Petersen, J. Berner, Mathematical capabilities of ChatGPT, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 27699–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/58168e8a92994655d6da3939e7cc0918-Paper-Datasets_and_Benchmarks.pdf.

[16] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, W. Yin, Large language models for mathematical reasoning: Progresses and challenges, in: N. Falk, S. Papi, M. Zhang (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, St. Julian's, Malta, 2024,

pp. 225–237. URL: https://aclanthology.org/2024.eacl-srw.17/.

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[18] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, I. Stoica, Chatbot arena: an open platform for evaluating LLMs by human preference, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.

[19] L. Zotos, H. van Rijn, M. Nissim, Are you doubtful? Oh, it might be difficult then! Exploring the use of model uncertainty for question difficulty estimation, in: C. Mills, G. Alexandron, D. Taibi, G. L. Bosco, L. Paquette (Eds.), Proceedings of the 18th International Conference on Educational Data Mining, International Educational Data Mining Society, Palermo, Italy, 2025, pp. 77–89. doi:10.5281/zenodo.15870153.

[20] P. Pezeshkpour, E. Hruschka, Large language models sensitivity to the order of options in multiple-choice questions, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2006–2017. URL: https://aclanthology.org/2024.findings-naacl.130/. doi:10.18653/v1/2024.findings-naacl.130.

[21] B. Plaut, K. Nguyen, T. Trinh, Softmax probabilities (mostly) predict large language model correctness on multiple-choice Q&A, CoRR abs/2402.13213 (2024). URL: https://doi.org/10.48550/arXiv.2402.13213.

[22] L. Zotos, H. van Rijn, M. Nissim, Can model uncertainty function as a proxy for multiple-choice question item difficulty?, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 11304–11316. URL: https://aclanthology.org/2025.coling-main.749/.