Are You Doubtful? Oh, It Might Be Difficult Then! Exploring the Use of Model Uncertainty for Question Difficulty Estimation

Anonymous ACL submission

Abstract

In an educational setting, an estimate of the difficulty of multiple-choice questions (MCQs), a commonly used strategy to assess learning progress, constitutes very useful information for both teachers and students. Since human assessment is costly from multiple points of view, automatic approaches to MCQ item difficulty estimation are investigated, yielding however mixed success until now. Our approach to this problem takes a different angle from previous work: asking various Large Language Models to tackle the questions included in two different MCQ datasets, we leverage *model uncertainty* to estimate item difficulty. By using both model uncertainty features as well as textual features in a Random Forest regressor, we show that uncertainty features contribute substantially to difficulty prediction, where difficulty is inversely proportional to the number of students who can correctly answer a question. In addition to showing the value of our approach, we also observe that our model achieves state-of-the-art results on the BEA publicly available dataset.

1 Introduction

Multiple-Choice Questions (MCQs) are commonly used as a form of assessment across educational levels. This is not surprising, as they are trivial to grade and can effectively assess a student's knowledge, as long as they are designed well (Gierl et al., 2017). Naturally, an aspect that significantly affects an MCQ's quality is its *difficulty*. Intuitively, items that are too easy do not sufficiently challenge students, while very difficult items lead to frustration and demotivation (Papoušek et al., 2016) impairing the learning process. However, estimating an item's difficulty is not trivial. In fact, students, and especially teachers, are not great at estimating how many of the test-takers will select the correct answer, given a question (van de Watering and van der Rijt, 2006). While field-testing question items is a viable solution, it is usually expensive, both in terms of time and resources.

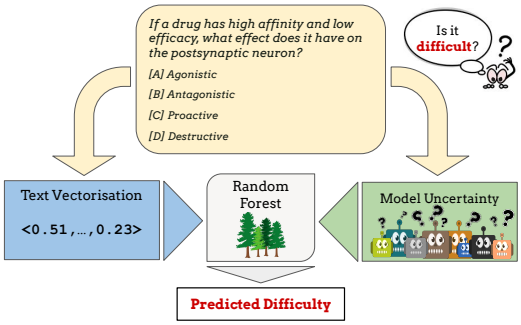


Figure 1: Approach overview: Predicting difficulty of Multiple-Choice Question items using textual features and uncertainty of LLM test-takers.

Computational methods, including Large Language Models (LLMs), have had some success in assessing the difficulty of MCQs (AlKhuzaey et al., 2024). At the same time, the task remains challenging, as shown by a recent shared task on automated difficulty prediction for MCQs (Yaneva et al., 2024), where most submitted systems performed barely above some simple baselines. The goal of the current work is to tackle the task of item difficulty estimation using a minimal experimental setup showcasing the usefulness of *model uncertainty* for this task. We do this by obtaining a score for the uncertainty LLMs exhibit when answering a variety of MCQs and use it, in combination with basic text and semantic features, to train an interpretable regressor model. This expands on previous findings which showed a correlation between model and student perceived difficulty (Zotos et al., 2024), paired with the intuition that both syntactic and semantic features are integral to this task (AlKhuzaey et al., 2024). We focus on factual MCQs, as this type of assessment is less subjective compared to open-ended questions, while still offering more complexity than simple True/False

questions[1]. Lastly, this choice is also motivated by dataset availability, as explained in Section 3.

It is worthwhile explicitly mentioning that in the current work, the term "uncertainty" is used to encompass both 1st token probability and choice-order probability metrics (see Section 4.2 for details.) These measures are taken to broadly represent the inverse of model confidence. While accurately determining the uncertainty of an LLM is an open field of research, previous research suggests that both 1st token probability (Plaut et al., 2024) and choice order probability (Zotos et al., 2024) correlate well with model correctness in the MCQs setup. These findings also hold in the current experimental setup, as shown in Appendix A.

Our Contribution The contribution of our work is twofold. First, thanks to extensive experiments with a variety of LLMs and feature analysis using an interpretable model (Random forest Regressor), we showcase that model uncertainty is a useful proxy for item difficulty estimation on two different question sets assessing factual knowledge. Second, as a byproduct of our experiments investigating model uncertainty we yield a model which achieves best results to date on the BEA 2024 Shared Task dataset. This model, together with all experimental code, is made available to the community for replicability and future extensions. We believe that our conceptual insight (model uncertainty as a useful signal for item difficulty), as well as our practical contribution in terms of an existing modular system, will foster further improvements in the task of MCQ automatic difficulty estimation, which is core in the educational setting.[2]

2 Related work

The task of estimating the difficulty of MCQ items has been explored from various viewpoints in the literature (AlKhuzaey et al., 2024). Most commonly this task is tackled by training a model on a set of syntactic (Perkins et al., 1995; Ha et al., 2019, e.g.,) and/or semantic features (Xue et al., 2020; Hsu et al., 2018, e.g.,). Furthermore, the majority of studies focus on the field of Language learning (Bi et al., 2021; He et al., 2021, e.g.,) which is inherently different to factual knowledge examinations. While the task of difficulty estimation has been widely explored, it remains challenging as was also seen in the recent "Building Educational Applications" (BEA) shared task on "Automated Prediction of Item Difficulty and Item Response Time", where simple baselines were overall only marginally beaten (Yaneva et al., 2024). In this task, a variety of approaches were explored with the focus ranging from architectural changes to data augmentation techniques. Notably, the best performing team (EduTec) used a combination of model optimisation techniques, namely scalar mixing, rational activation and multi-task learning (leveraging the provided response time measurements also provided in the BEA dataset) (Gombert et al., 2024).

Most similar to our work is the study by Loginova et al. (2021), who also explore the use of confidence of language models to estimate question difficulty. While similarities exist, the current research deviates considerably from this study. Their focus is on language comprehension, which differs from our emphasis on factual knowledge assessment. Furthermore, while their focus is on Encoder-Only models, ours is on Decoder-Only models, which incorporate greater amounts of factual knowledge as a byproduct of their language modeling objective (Zhao et al., 2023).

More broadly, there is an emerging "LLM-as-a-judge" field of research, which, in general terms, explores the possibility of using powerful LLMs as a substitute for human annotation (Zheng et al., 2023; Pan et al., 2024). For the task of question difficulty estimation, this paradigm has been explored in the context of language comprehension by Raina and Gales (2024) with some success. However, our preliminary research suggests that this approach does not perform well for the two datasets used in the current study (see Appendix B), thereby motivating further exploration in novel approaches.

The present work directly builds on the work by Zotos et al. (2024), where a variety of analyses showed promising results on correlating human and machine perceived difficulty. We take this one step further, by testing a battery of different LLMs on item difficulty estimation using their uncertainty as a signal, focusing on two distinct question sets assessing factual knowledge.

3 Data

The two factual knowledge MCQ datasets that we use in our experiments are described more in detail in the following subsections. The first is a dataset

[1] In True/False questions, a statement needs to be assessed as correct or incorrect, with a random chance of 50%. In contrast, MCQs can follow various formulations and the distractors play a significant role.

[2] Code will become available upon acceptance.

on the domain of Biopsychology that is not publicly available, while the second is the publicly available dataset used in the BEA 2024 Shared Task (Yaneva et al., 2024). For brevity, we refer to the "Biopsychology" and "BEA" datasets respectively. Our choice is driven by the requirement of having question-sets along with students selection rates (serving as proxies for item difficulty scores). Considering that, to the best of our knowledge, the BEA dataset is the only publicly available resource satisfying this requirement, we also use a non-publicly available dataset. This choice is in line with the observation by AlKhuzaey et al. (2024), who note that most studies tackling this task resort to using private datasets.

We use first the Biopsychology dataset to extend the experiments in (Zotos et al., 2024), and then the BEA dataset to evaluate whether our approach generalises to a different dataset (for which many other systems exist and can be compared to). As will be explained in Sections 3.1 and 3.2, even though both question sets assess factual knowledge, they also vary in multiple aspects, for example question formulation, number of distractors and knowledge specificity. Furthermore, to facilitate comparison with the findings from the BEA 2024 Shared Task, we use the train/test split as provided in the shared task itself (70% training and 30% test samples). The same proportions are also used for the Biopsychology dataset, as shown in Table 1.

The item difficulty labels differ between the two datasets. In the Biopsychology dataset, difficulty is measured by the proportion of students who answered correctly (a higher value indicates an easier question). In contrast, the BEA dataset originally uses the inverse difficulty measurement, where a higher difficulty label signifies that fewer students answered correctly. Additionally, a linear transformation is also applied on the target labels of the BEA dataset. While this difference does not affect our approach, as in both cases difficulty is conceptually expressed by cumulative student performance, to allow easier interpretation of our results we have transformed the BEA difficulty scores to their complements such that they also reflect the proportion of correct responses per question.

3.1 Biopsychology

The Biopsychology dataset originates from a course taught in the 1st year of the Psychology undergraduate degree at a Social Sciences Faculty, covering content from the classic textbook "Biological Psychology" by Kalat (2016). The dataset comprises of 819 MCQs in total, of which 451 and 368 have two and three distractors respectively. The data was collected from fifteen examinations with an average of 261 examinees (Standard Deviation of 184). This dataset has not been previously made public, minimising the risk of data contamination (ensuring that the LLMs used have not encountered the question set during training). An important feature of this question set is its high textual variability, with questions ranging from "Fill two gaps" to "Wh-questions". Two example questions are reported in Table 2. Given that LLMs demonstrate sensitivity to input formulation (Biderman et al., 2024), the presence of such variability in the data improves generalisation of our method across datasets.

Dataset	Train	Test	Total
Biopsychology	573	246	819
BEA	466	201	667

Table 1: Train and Test splits as used in our experiments. For BEA, we use the splits as provided in the competition (Yaneva et al., 2024). For Biopsychology, we randomly sampled the questions, keeping the same percentage of training/testing samples as in BEA.

3.2 BEA 2024

The BEA question set was used in the United States Medical Licensing Examination (USMLE®) and was developed by the National Board of Medical Examiners (NBME®) and Federation of State Medical Boards (FSMB) (Yaneva et al., 2024). It consists of 667 MCQs, each answered by more than 300 medical school students. In contrast to the Biopsychology dataset, the questions follow strict guidelines (e.g., fixed structure, absence of misleading or redundant information in the question) and are presented with up to nine distractor choices, with the majority of the questions having five (525 items) or six distractors (71 items). An example instance is provided in Table 2. As can be seen, questions of this dataset are significantly longer (755 characters compared to 103 characters for the Biopsychology set) and are of technical nature.

4 Approach

Given an MCQ, the task is to predict the proportion of students that select the correct choice[3]. An

[3]This is also known as the *p-value* (van de Watering and van der Rijt, 2006).

Dataset	Question	Choices
Biopsychology	Homeostasis is to ... as allostasis is to ...	ⓐ constant; variable ⓑ constant; decreasing ⓒ variable; constant
Biopsychology	If a drug has high affinity and low efficacy, what effect does it have on the postsynaptic neuron?	ⓐ agonistic ⓑ antagonistic ⓒ proactive ⓓ destructive
BEA	A 65-year-old woman comes to the physician for a follow-up examination after blood pressure measurements were 175/105 mm Hg and 185/110 mm Hg 1 and 3 weeks ago, respectively. She has well-controlled type 2 diabetes mellitus. Her blood pressure now is 175/110 mm Hg. Physical examination shows no other abnormalities. Antihypertensive therapy is started, but her blood pressure remains elevated at her next visit 3 weeks later. Laboratory studies show increased plasma renin activity; the erythrocyte sedimentation rate and serum electrolytes are within the reference ranges. Angiography shows a high-grade stenosis of the proximal right renal artery; the left renal artery appears normal. Which of the following is the most likely diagnosis?	ⓐ Atherosclerosis ⓑ Congenital renal artery hypoplasia ⓒ Fibromuscular dysplasia ⓓ Takayasu arteritis ⓔ Temporal arteritis

Table 2: Examples questions from the Biopsychology and BEA datasets. Correct answer in green.

MCQ item consists of the stem/question, a single correct choice/answer and a number of incorrect choices/distractors (also known as “foils").

Figure 1 illustrates our approach to this task. Our design is centered around a simple Random Forest Regressor[4] which receives as input a vectorised representation of the MCQ, as well the uncertainty of multiple LLMs answering the same MCQ[5]. We opted for a relatively simple Random Forest Regressor due to its interpretability compared to more complex architectures, while still effectively demonstrating the usefulness of model uncertainty in this context. As features, we use Textual Features and Model Uncertainty, as described in the following sections.

4.1 Textual Features

Intuitively, extracting the semantic content of the question item is integral to assess its difficulty. To accomplish that, we use two fundamentally different methods – Term Frequency-Inverse Document Frequency (TF-IDF) Scores and Semantic Embeddings – to encode the question and answer choices as numerical vectors.

TF-IDF Scores TF-IDF Scores capture how important a word is to a document within a collection by balancing its frequency in that specific document against its rarity across all documents (Sparck Jones, 1972). In the current context, we consider each question item (along with its choices) as a single document. To capture multi-word technical terms, such as "interstitial fibrosis", our analysis considers both individual words (unigrams) and two-word combinations (bigrams). Furthermore, we disregard terms that appear in more than 75% of documents, and only use the 1000 most important features (as determined by the TF-IDF values) to increase efficiency.

Semantic Embeddings Word embeddings are a technique whereby words are encoded as dense vectors in a continuous vector space, capturing semantic relationships between words. We evaluate two embedding approaches: General BERT embeddings (Devlin et al., 2019) and domain-specific Bio-Clinical BERT embeddings (Alsentzer et al., 2019), relevant to the topics addressed in the MCQ datasets we use. The Bio-Clinical BERT embeddings, previously also employed by team ITEC in the BEA 2024 shared task (Tack et al., 2024), offer specialized medical domain text encoding that potentially encapsulates more accurately the semantic content of each question item. Both techniques yield a 768-dimensional vector representation.

4.2 Model Uncertainty

The current methodological approach is founded on the premise that model uncertainty correlates with student performance and thus, by extension, offers a useful signal when estimating the difficulty of a question item. To explore this hypothesis, we have conducted experiments using two metrics that are

[4]As provided by the Scikit-Learn Library, using the default hyper-parameters (Pedregosa et al., 2011).

[5]Simple vector concatenation is used to combine the text and uncertainty features.

shown to correlate well with model correctness (as discussed in Appendix A): *1st Token Probability* and *Choice Order Sensitivity*. These uncertainty scores are obtained for each LLM separately and concatenated into a single vector, to which textual features are (optionally) also added. This vector is then fed to the regressor.

1st Token Probability The first technique to measure model uncertainty is by inspecting the softmax probability of the 1st token to be generated as the answer id to the given MCQ question, (e.g., probability of generating token "B"), in comparison to the probabilities of the alternatives (e.g., probability of generating token "A" or "C"). As the 1st token probabilities can be influenced by the order in which the choices are provided in the problem set (Wei et al., 2024; Wang et al., 2023, 2024; Zheng et al., 2024), we create ten random different orderings for each question and let the model answer each MCQ ten times[6]. This way, we calculate the average probability per MCQ choice. We then consider the average probability for the correct answer as the uncertainty metric of the LLM.

Furthermore, as different tokens might be generated to represent the same answer (e.g., "A", " A", "a ", see details on prompting and answer elicitation in Section 4.3 below) and different models might attribute higher likelihood for specific tokens, the token representing each choice with the highest probability is selected. For example if for a given model the probability of generating token "C" is higher than the probability of token "c", the former is considered for that model. Lastly, the three extracted mean probabilities of all orderings are normalised in the range of $0-1$ such that they can be more easily compared to the difficulty scores, which – being calculated as proportions of the student populations – are in the same range.

Choice Order Sensitivity Pezeshkpour and Hruschka (2023) observed that choice order sensitivity correlates with error rate. In other words, when LLMs consistently select a choice regardless of its position, that choice is more likely to be correct. Based on this observation, we leverage this correlation to measure uncertainty. Specifically, for all evaluated choice orderings, we measure the probability of the correct choice being selected. Thus, this probability is not based on token probabilities but rather on the eventual choice.

[6]For questions with only 3 choices, we instead consider all six different choice orderings.

Instruction Prompt for the LLM

Below is a multiple-choice question. Choose the letter which best answers the question. Keep your response as brief as possible; just state the letter corresponding to your answer with no explanation.
Question:
[Question Text]
Response:

Figure 2: Instruction phrasing used for all models and experiments. *[Question Text]* is replaced by the item stem followed by the answer choices, each prepended with the corresponding letter *A* to *J*.

4.3 Choice of Models and Prompting

In this work, we focus on decoder-only models, as they are considered to have incorporated greater amounts of factual knowledge as a byproduct of their language modelling objective (Zhao et al., 2023), compared to Encoder-Only or Encoder-Decoder models. Moreover, as the internal logit probabilities of the 1st token to be generated are needed to measure the uncertainty of each model, we focus on nine open-sourced models of different parameter sizes and families. Additionally, we constrict our choice to instruction-tuned models and use 4-bit quantisation for increased efficiency.[7] To adapt them for the task of MCQ answering, we use the instruction prompt in Figure 2 based on (Plaut et al., 2024) and (Zotos et al., 2024).

5 Results

Our experiments are aimed at evaluating the usefulness of model uncertainty as a signal for MCQ item difficulty as well as discovering which specific textual and uncertainty features are most relevant for our trained Regressor. We first focus on the performance of our setup using different feature sets (Section 5.1), followed by an in-depth analysis of the importance of individual features (Section 5.2).

5.1 Performance on Difficulty Estimation

To evaluate the performance of our trained models we use the Root Mean Squared Error (RMSE) metric from Python's Scikit-learn library (Pedregosa et al., 2011), as used in the BEA 2024 Shared task. As previously mentioned, we use a Random Forest

[7]Details of the models used are in Table 6 in Appendix C.

	Biopsychology		BEA	
Dummy Regressor	0.1670		0.3110	
Best BEA 2024 Competition Result (Team EduTec)	-		0.2990	
Only Text Features				
TF-IDF	0.1479		0.3092	
BERT Embeddings	0.1498		0.3066	
Only Model Uncertainty Features				
1st Token Probabilities	0.1539		0.2960	
Choice Order Sensitivity	0.1582		0.3178	
1st Token Probability & Choice Order Sensitivity	0.1538		0.2968	
Text and Model Uncertainty Features				
	TF-IDF	**BERT**	**TF-IDF**	**BERT**
First Token Probability	0.1365	0.1385	0.2851	0.2854
Choice Order Sensitivity	**0.1309**	0.1411	0.2951	0.2961
1st Token Probability & Choice Order Sensitivity	0.1371	0.1388	0.2856	**0.2846**

Table 3: Root Mean Squared Error (RMSE, the lower the better) on the test set using different sets of features. Lowest achieved RMSE per dataset is shown in **boldface**. All results are averaged over ten repetitions, with the standard deviation not exceeding 0.002.

Regressor tasked to predict the difficulty of a question item, given as input a vectorised representation of the MCQ as well as the uncertainty of multiple LLMs answering the same MCQ. This creates a modular setup that allows easy manipulation of the input feature set. We present the feature sets along with their performance on the two datasets in Table 3. For brevity, we report the results obtained using Bio-Clinical BERT Embeddings in Appendix D, as they were found to lead to similar, yet consistently slightly worse RMSE scores compared to the general BERT embeddings.

An important first observation is that the RMSE difference between experiments is minimal. This is in-line with the findings from the BEA 2024 shared task, where the lowest achieved RMSE was only 0.012 lower than the baseline. However, even though the margins are narrow, there are consistent differences between the experimental setups. Most importantly for this research, it is clear that using the uncertainty of the models, combined with text features yields significantly lower RMSE for both datasets, even beating the best score achieved during the BEA competition by a margin of 0.0125. In fact, providing the Random Forest Regressor only with LLMs' uncertainties also surpasses the best BEA competition result, albeit by a narrower margin of 0.003, further underscoring the potential of using model uncertainty for this task. Furthermore, in our exploration we did not find a consistent superiority for one of the two model uncertainty metrics (1st Token Probability or Choice Order Sensitivity) or text vectorisation methods (TF-IDF or BERT embeddings), though there seemed to be a general advantage in using TF-IDF scores over BERT Embeddings. Still, the observed scores for this subset of results are within the observed standard deviation of 0.002 between repetitions. Lastly, the usefulness of the model uncertainty features is especially clear in the experiments where the Random Regressor did not use any text features: except where only Choice Order Sensitivity is used in the BEA dataset, the performance is consistently better than the respective Mean Regressor baselines.

5.2 Importance of Different Features

In order to better understand which features drive the predictions of the Random Forest Regressor, we use Shapley additive explanations as provided by the SHAP Python library (Lundberg and Lee, 2017). To maintain conciseness, we present SHAP summary plots for a selected subset of experiments that we found to be the most insightful. Additional analyses are presented in Appendix E.

Before exploring the analysis regarding model uncertainty, we examine the contribution of the most impactful uni/bi-grams from the text-only experiment using the Biopsychology dataset. This analysis relies on TF-IDF scores, as BERT embed-

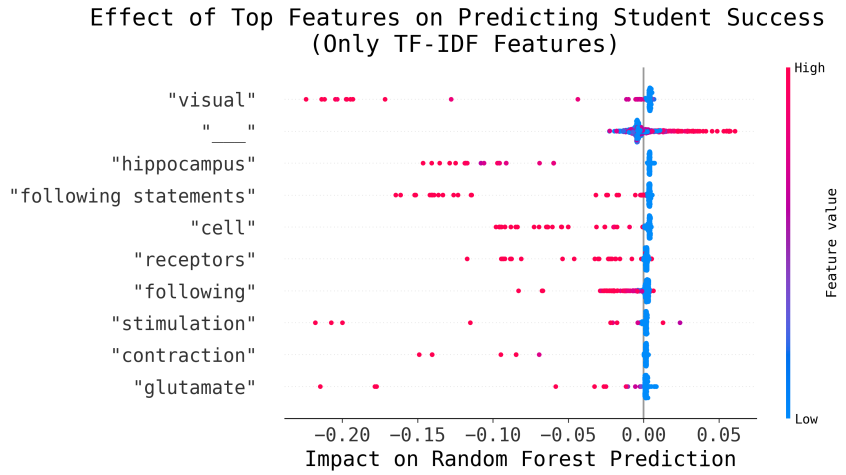


Figure 3: Biopsychology Dataset. SHAP summary plot showing the contribution of the top ten uni/bi-gram features to the Random Forest's predictions, highlighting their importance and impact direction. Features are ranked by their average influence, with dots representing individual question items and colour indicating TF-IDF scores. Results averaged over ten repetitions.

dings cannot directly be traced back to individual words. Figure 3 shows the ten most impactful features, along with their effect on the Regressors' prediction for each MCQ item. High TF-IDF scores, highlighted in red, indicate that an MCQ item is predicted to be more difficult (i.e., fewer students answer it correctly). For instance, questions containing the unigram "visual" prompted the Random Forest model to predict greater difficulty. Interestingly, this analysis also demonstrates that questions where a gap (represented by an underscore "_") needs to be filled (e.g., "fill-the-gap" or sentence completion) are predicted to be easier.

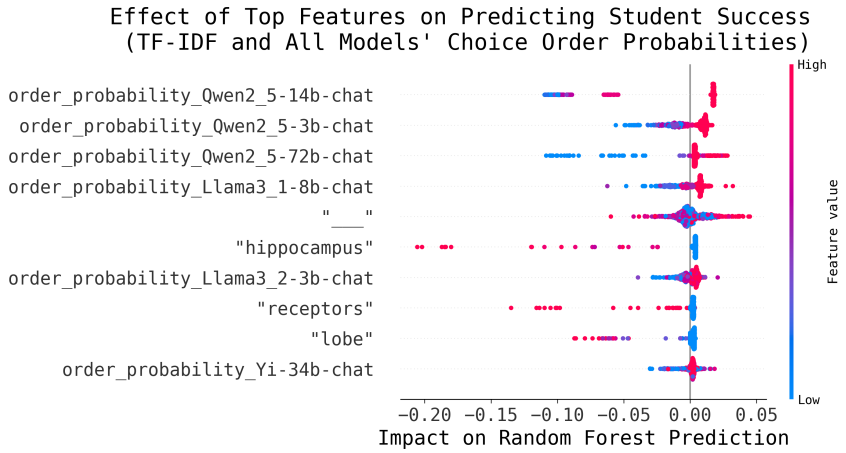
While this analysis gives some insights on important text features, we are mostly interested in the contribution of features related to model uncertainty. Figures 4a and 4b present the effect of the most impactful features for the feature sets that lead to the best performance (using TF-IDF scores for the text encoding and model uncertainty) for the Biopsychology and BEA datasets, respectively. In both instances, the Random Forest Regressor heavily relies on model uncertainties to predict item difficulty. As hypothesised, the higher the model certainty (in terms of either 1st Token or Choice-Order Probability) the more students are predicted to answer the question correctly. In terms of models, we observe that the confidence of Qwen2.5-14B-chat is especially useful for both datasets (as can be seen by its 1st and 2nd place in Figures 4a and 4b respectively). This observation also highlights the core challenge of our approach: having a model that is sufficiently capable of answering the MCQs but not so complex that it answers them with complete confidence. In our work, this challenge is partially addressed using an ensemble of models, leaving it up to the Random Forest Regressor to determine their usefulness.

6 Discussion and Conclusion

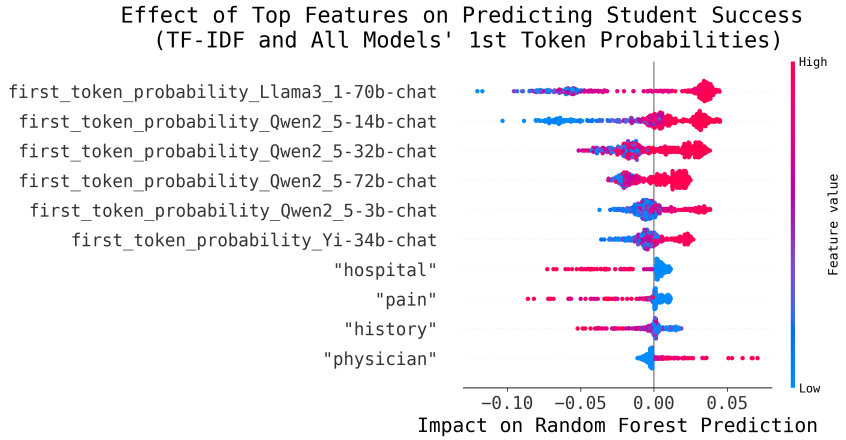
We explored how model uncertainty can be used as a proxy for item-difficulty using two MCQ datasets focusing on factual knowledge. We demonstrate, in a simple experimental setup, that while both textual features (e.g., encoding using TF-IDF Scores or BERT embeddings) and model uncertainty features are useful for the task, the trained Random Forest Regressor performed significantly better when model uncertainty features were included.

Our results suggest that aspects of a question item that challenge students similarly impact LLMs. A factor that could explain this alignment is representation: Knowledge that is well represented in an LLM's training data is likely to be more foundational (e.g., "What is a neuron"), compared to specialised knowledge (e.g., a medical diagnosis). By extension, using model uncertainty for this task requires a model of appropriate size/capabilities.

Our methodological design is intentionally simple, serving as a proof of concept for this approach. This simplicity stems from various design choices. Firstly, our definition of item difficulty is simply the number of students answering an MCQ correctly. This is in contrast to finer-grained metrics



(a) Biopsychology Dataset



(b) BEA Dataset

Figure 4: SHAP summary plots for the Biopsychology and BEA datasets showing the contribution of the top ten features to the Random Forest's predictions. Higher First Token Probability and Order Probability metrics indicate greater model certainty. Results averaged over ten repetitions.

such as Item Response Theory (Lord and Novick, 2008) which however require individual students' responses that are unavailable in our datasets. Secondly, we use a variety of LLMs without placing great emphasis on their uncertainty behaviour. Specifically, while we ensure that the measured model uncertainty aligns with model correctness (as shown in Appendix A), we do not focus on calibrating the LLMs. Instead, we rely on the Random Forest Regressor to select and weight the uncertainties of the various models. Thirdly, we limited this research to only the necessary features for the purpose of our study (assessing the contribution of model uncertainty), namely text and model uncertainty. In order for yet further improvements to be obtained on the actual task of difficulty estimation, more complex features can be explored and incorporate. For example, building on the intuition that the nature of the distractors plays a role in the question's difficulty, we experimented with a choice similarity metric, defined (per MCQ) as the average cosine similarity between each distractor and the correct answer choice, similar to the approach by (Susanti et al., 2020). Incorporating this information in our model yields an RMSE of 0.2841 on the BEA dataset, the lowest observed sofar. This improvement is only observed when this feature is combined with model uncertainty.[8] Shifting into a broader perspective, our findings incentivise further research into better understanding which methods capturing model uncertainty, ranging from analysing model internals to studying model output (e.g., related to prompt sensitivity) are most beneficial for this task.

[8]Experimental details and full results for this additional feature are presented in Table 8 in Appendix F.

Limitations

Indisputably, the central limitation of our approach is the reliance on (un)certain LLMs. As seen in Section 5, model uncertainty is beneficial only when the model can answer the question without being overly confident. As a result, limits the usefulness of our approach, especially given the rapid development of LLMs in terms of their capabilities. We hypothesise that this limitation can at least partially be resolved by using calibrated LLMs, which we leave for future work.

Similarly, our approach is not expected to perform as well on MCQs designed to test knowledge at lower education levels (e.g., primary school geography exams), as even small LLMs are now capable of confidently answering such questions. At the same time, using less proficient LLMs introduces different challenges, particularly regarding linguistic ability: Smaller LLMs are more strongly affected by linguistic perturbations (e.g., question formulation, choice order) and have greater limitations in instruction-following capabilities (Biderman et al., 2024; Sclar et al., 2023).

Lastly, due to dataset availability, we evaluated our approach solely on factual knowledge examinations. It remains unclear whether model uncertainty could also be beneficial for assessing the difficulty of examinations of other skill sets, such as language comprehension or mathematical reasoning.

Ethics Statement

In this study, we used a dataset of multiple-choice questions from the "Biopsychology" course at the Behavioural and Social Sciences Faculty of our institution. The data was aggregated across multiple students and anonymised, ensuring that individual student performance cannot be traced.

References

Samah AlKhuzaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Lizhen Qu, Shirong Shen, Guilin Qi, Lu Pan, and Yinlin Jiang. 2021. Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4645–4654, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the Trenches on Reproducible Evaluation of Language Models. *Preprint*, arXiv:2405.14782.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang. 2017. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87:0034654317726529.

Sebastian Gombert, Lukas Menzel, Daniele Di Mitri, and Hendrik Drachsler. 2024. Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 483–492, Mexico City, Mexico. Association for Computational Linguistics.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 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, Florence, Italy. Association for Computational Linguistics.

Jun He, Li Peng, Bo Sun, Lejun Yu, and Yinghui Zhang. 2021. Automatically predict question difficulty for reading comprehension exercises. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1398–1402. IEEE.

Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing and Management*, 54:969–984.

James W. Kalat. 2016. *Biological psychology*. Cengage Learning.

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, Held Online. INCOMA Ltd.

Frederic M Lord and Melvin R Novick. 2008. *Statistical theories of mental test scores*. IAP.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Qian Pan, Zahra Ashktorab, Michael Desmond, Martin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-centered design recommendations for llm-as-a-judge. *Preprint*, arXiv:2407.03479.

Jan Papoušek, Vít Stanislav, and Radek Pelánek. 2016. Impact of question difficulty on engagement and learning. In *Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings 13*, pages 267–272. Springer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kyle Perkins, Lalit Gupta, and Ravi Tammana. 1995. Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language Testing*, 12(1):34–53.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *Preprint*, arXiv:2308.11483.

Benjamin Plaut, Khanh Nguyen, and Tu Trinh. 2024. Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a. *Preprint*, arXiv:2402.13213.

Vatsal Raina and Mark Gales. 2024. Question Difficulty Ranking for Multiple-Choice Reading Comprehension. *Preprint*, arXiv:2404.10704.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Yuni Susanti, Takenobu Tokunaga, and Hitoshi Nishikawa. 2020. Integrating automatic question generation with computerised adaptive test. *Research and Practice in Technology Enhanced Learning*, 15:1–22.

Anaïs Tack, Siem Buseyne, Changsheng Chen, Robbe D'hondt, Michiel De Vrindt, Alireza Gharahighehi, Sameh Metwaly, Felipe Kenji Nakano, and Ann-Sophie Noreillie. 2024. ITEC at BEA 2024 shared task: Predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 512–521, Mexico City, Mexico. Association for Computational Linguistics.

Gerard van de Watering and Janine van der Rijt. 2006. 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(2):133–147.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. 2024. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. *Preprint*, arXiv:2404.08382.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. *Preprint*, arXiv:2406.03009.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197, Seattle, WA, USA → Online. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. 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), pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. 2024. Can model uncertainty function as a proxy for multiple-choice question item difficulty? *Preprint*, arXiv:2407.05327.

A Model Correctness and Uncertainty

Table 4 presents the performance of each model on the two question sets, as well as the relation between model certainty and model correctness. In line with the results of Plaut et al. 2024, it is clear that both tested metrics correlate well with model correctness: For all models, on average, the mean certainty is significantly higher for the correctly answered question items. This suggests that the two metrics indeed capture an aspect of model certainty.

B LLM-as-a-Judge Approach

In this section we briefly explore the possibility of using a strong LLM, Llama3.1-70B-Chat, to tackle the task of MCQ item difficulty estimation. Specifically, for each of the two datasets we instruct the LLM to predict the number of students correctly answering the given question, using the prompt presented in table 5. This approach results in RMSEs of 0.2881 and 0.3565 for the Biopsychology and BEA datasets respectively, which is significantly worse than the mean regressor baseline.

C Large Language Models Used

Table 6 presents the collection of Large Language Models used in our experiments involving model uncertainty.

D Use of Bio_Clinical Bert Embeddings

Table 7 presents the performance of our setup using the Bio_clinical Bert Embeddings (Alsentzer et al., 2019). The results using BERT embeddings (Devlin et al., 2019) are repeated to facilitate comparison. In our experiments, using the Bio_clinical Bert Embeddings consistently led to worse performance (higher RMSE) for all experiments.

E Additional SHAP analyses

In this section we present a collection of additional analyses using the Shapley additive explanations (Lundberg and Lee, 2017).

E.1 Only Using TF-IDF Features using BEA Dataset

Figure 5 presents the effect of the ten most impactful uni/bi-grams on the trained Random Forest's difficulty prediction. Questions containing the words "physician" and "examination" are generally predicted to be easy, while the other impactful features lead to the Random Forest predicting higher difficulty (lower proportion of students selecting the correct answer).

E.2 TF-IDF Features and All Model Uncertainty Features

Figure 6 shows the SHAP summary plot when both model uncertainty feature-sets are given to the Random Forest Regressor, along with the text encoded as TF-IDF scores. Here, we observe that overall 1st Token Probabilities are preferred over Choice-Order Probabilities as a proxies for difficulty.

F Using Semantic Similarity Between Answer Choices

While the central goal of our study is to showcase that model uncertainty is a useful signal for MCQ item difficulty, in this section, we show that additional features have the potential to further the performance of our setup. More in detail, we use a choices similarity metric, defined (per MCQ) as the average cosine similarity between each distractor and the correct answer choice, similar to the approach by (Susanti et al., 2020). This is operationalised using the Sentence Transformer library (Reimers and Gurevych, 2019) with one of two models: "all-MiniLM-L6-v2"[9] (general efficient

[9]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Dataset	Model	Overall Correctness	Mean Probability	
			1st Token	Choice Order
Biopsychology	phi3_5-chat	0.302	0.448 / 0.232	0.532 / 0.240
	Llama3_2-3b-chat	0.707	0.733 / 0.247	0.853 / 0.196
	Qwen2_5-3b-chat	0.799	0.864 / 0.200	0.883 / 0.192
	Llama3_1-8b-chat	0.824	0.683 / 0.266	0.860 / 0.236
	Qwen2_5-14b-chat	0.902	0.972 / 0.136	0.975 / 0.115
	Qwen2_5-32b-chat	0.932	0.968 / 0.154	0.978 / 0.121
	Yi-34b-chat	0.852	0.878 / 0.234	0.907 / 0.215
	Qwen2_5-72b-chat	0.937	0.964 / 0.175	0.984 / 0.139
	Llama3_1-70b-chat	0.933	0.946 / 0.209	0.980 / 0.160
BEA	phi3_5-chat	0.193	0.368 / 0.158	0.426 / 0.152
	Llama3_2-3b-chat	0.645	0.607 / 0.183	0.782 / 0.153
	Qwen2_5-3b-chat	0.510	0.676 / 0.148	0.752 / 0.123
	Llama3_1-8b-chat	0.654	0.441 / 0.194	0.758 / 0.169
	Qwen2_5-14b-chat	0.741	0.905 / 0.163	0.915 / 0.150
	Qwen2_5-32b-chat	0.805	0.912 / 0.170	0.936 / 0.151
	Yi-34b-chat	0.651	0.748 / 0.175	0.802 / 0.155
	Qwen2_5-72b-chat	0.855	0.916 / 0.195	0.953 / 0.169
	Llama3_1-70b-chat	0.892	0.845 / 0.181	0.945 / 0.144

Table 4: Model correctness and answer probability in terms of Mean 1st Token and Choice Order Probability in the Biopsychology and BEA question sets. “Overall Correctness" indicates the proportion of correctly answered questions, while the probabilities in green and red indicate the mean model certainty for correctly and incorrectly answered questions respectively. As can be seen, on average, model certainty is significantly higher for questions that are answered correctly.

Instruction Prompt
Below is a multiple-choice question. Out of 100 students, how many do you think answered correctly? Answer with a number between 0 and 100 and do not include an explanation or any other text. Question: *[Question Text]* Number of students (out of 100) answering correctly:

Table 5: Instruction phrasing used for the LLM-as-a-Judge exploratory experiment. *[Question Text]* is replaced by the item stem followed by the answer choices, each prepended with the corresponding letter *A* to *J*.

Model	Source
phi3_5-chat	https://huggingface.co/unsloth/Phi-3.5-mini-instruct-bnb-4bit
Llama3_2-3b-chat	https://huggingface.co/unsloth/Llama-3.2-3B-Instruct-bnb-4bit
Qwen2_5-3b-chat	https://huggingface.co/unsloth/Qwen2.5-3B-Instruct-bnb-4bit
Llama3_1-8b-chat	https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit
Qwen2_5-32b-chat	https://huggingface.co/unsloth/Qwen2.5-32B-Instruct-bnb-4bit
Qwen2_5-14b-chat	https://huggingface.co/unsloth/Qwen2.5-14B-Instruct-bnb-4bit
Yi-34b-chat	https://huggingface.co/unsloth/yi-34b-chat-bnb-4bit
Qwen2_5-72b-chat	https://huggingface.co/unsloth/Qwen2.5-72B-Instruct-bnb-4bit
Llama3_1-70b-chat	https://huggingface.co/unsloth/Meta-Llama-3.1-70B-Instruct-bnb-4bit

Table 6: LLMs used in the experiments

	Biopsychology		BEA	
Only Text Features				
Bio_Clinical BERT	0.1518		0.3094	
BERT	0.1498		0.3066	
Text and Model Uncertainty Features				
	Bio_Clinical BERT	**BERT**	**Bio_Clinical BERT**	**BERT**
First Token Probability	0.1396	0.1385	0.2908	0.2854
Choice Order Sensitivity	0.1426	0.1411	0.3035	0.2961
1st Token Probability & Choice Order Sensitivity	0.1392	0.1388	0.2909	0.2846

Table 7: Comparison of results using Bio_Clinical BERT or default BERT embeddings in terms of RMSE on the test set. All results are averaged over ten repetitions, with the standard deviation not exceeding 0.002.

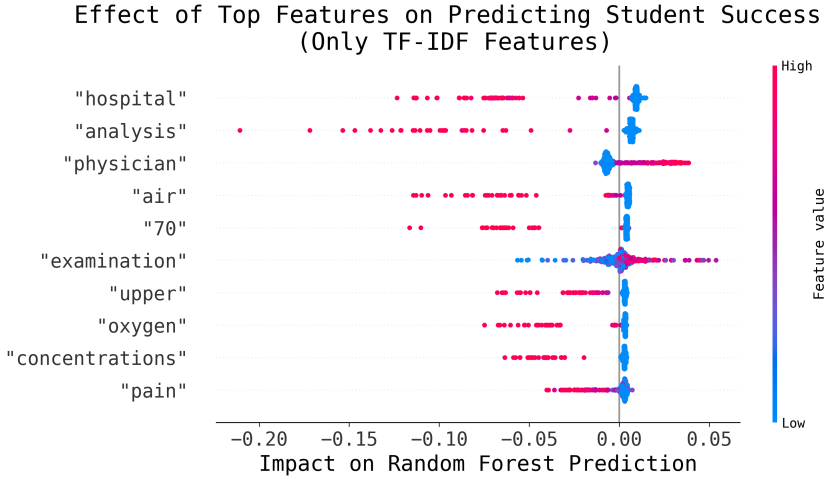
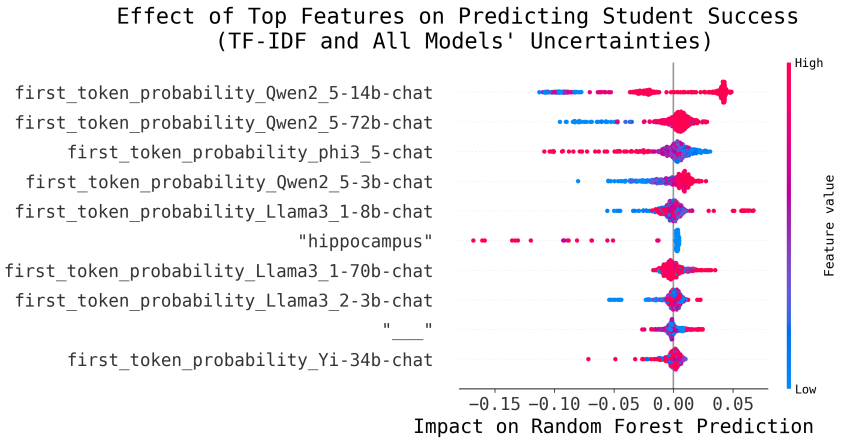


Figure 5: BEA Dataset. SHAP summary plot showing the contribution of the top ten uni/bi-gram features to the Random Forest's predictions, highlighting their importance and impact direction. Features are ranked by their average influence, with dots representing individual question items and colour indicating TF-IDF scores. Results averaged over ten repetitions.

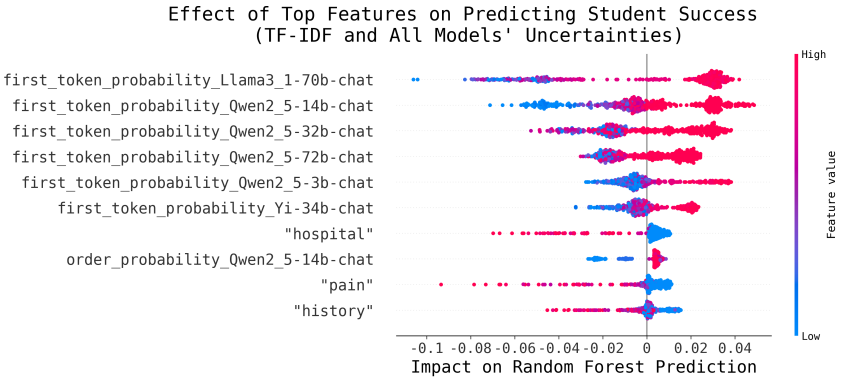
embeddings) and "S-PubMedBert-MS-MARCO"[10] (medical/health text domain embeddings). The two setups are henceforth refered to as "General Similarity" and "Medical Similarity" respectively.

Table 8 presents the achieved RMSE using text features, model uncertainty and choice similarity in the two datasets.

[10] pritamdeka/S-PubMedBert-MS-MARCO



(a) Biopsychology Dataset



(b) BEA Dataset

Figure 6: SHAP summary plots for the Biopsychology and BEA datasets showing the contribution of the top ten features to the Random Forest's predictions provided TF-IDF scores and all model uncertainty features. Higher First Token Probability and Order Probability metrics indicate greater model certainty. Results averaged over ten repetitions.

	RMSE			
Best Result without Choice Similarity				
Biopsychology: TF-IDF & Choice-Order Sensitivity	**0.1309**			
BEA: BERT Embeddings, 1st Token Probability & Choice-Order Sensitivity	0.2846			
	Biopsychology		**BEA**	
Only Choice Similarity Features				
General Similarity	0.1895		0.3567	
Medical Similarity	0.1883		0.3432	
Text, Model Uncertainty and Choice Similarity Features				
	TF-IDF	**BERT**	**TF-IDF**	**BERT**
1st Token Probability, Choice Order Sensitivity & General Similarity	0.1386	0.1389	0.2850	**0.2841**
1st Token Probability, Choice Order Sensitivity & Medical Similarity	0.1378	0.1381	0.2856	0.2844

Table 8: Performance of Random Forest Regressor using text, model uncertainty and choice similarity features. Best score per dataset indicated in **boldface**. All results are averaged over ten repetitions, with the standard deviation not exceeding 0.002.