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

Leonidas Zotos
Center for Language and Cognition
Groningen, The Netherlands
l.zotos@rug.nl

Hedderik van Rijn
Department of Experimental Psychology
Groningen, The Netherlands
d.h.van.rijn@rug.nl

Malvina Nissim
Center for Language and Cognition
Groningen, The Netherlands
m.nissim@rug.nl

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 three 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 USMLE and CMCQRD publicly available datasets.

Keywords

item difficulty estimation, model uncertainty, multiple choice questions

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 [6]. Naturally, an aspect that significantly affects an MCQ's quality is its *difficulty* (broadly determined by the students' success answering the question item). Intuitively, items that are too easy do not sufficiently challenge students, while very difficult items lead to frustration and demotivation impairing the learning process [17]. 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 [30]. While field-testing question items is a viable solution, it is usually expensive, both in terms of time and resources.

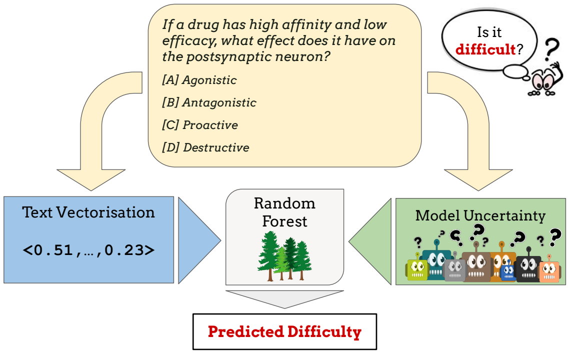


Figure 1: Approach overview: Predicting difficulty of Multiple-Choice Question items using textual features and uncertainty of LLM test-takers.

Computational methods, including Large Language Models (LLMs), have had some success in assessing the difficulty of MCQs [1]. At the same time, the task remains challenging, as shown by a recent shared task on automated difficulty prediction for MCQs [36], 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 a regressor model. This expands on previous findings which showed a correlation between model and student perceived difficulty [40], paired with the intuition that both syntactic and semantic features are integral to this task [1]. We focus on factual MCQs, as they provide more objective assessment than open-ended questions, while still offering more complexity than True/False questions where the a baseline random chance is 50%. In contrast, MCQs can follow various

Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. Are You Doubtful? Oh, It Might Be Difficult Then! Exploring the Use of Model Uncertainty for Question Difficulty Estimation. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 77–89. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870153

formulations and the incorrect choices play a significant role. 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 [22] and Choice-Order Probability [40] correlate well with model correctness in the MCQ 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 a regressor model (Random Forest Regressor), we showcase that model uncertainty is a useful proxy for item difficulty estimation on three different question sets assessing both factual knowledge and reading comprehension. 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 as well as the CMCQRD 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.[1]

2. RELATED WORK

The task of estimating the difficulty of MCQ items has been explored from various viewpoints in the literature [1]. Most commonly this task is tackled by training a model on a set of syntactic [20, 8] and/or semantic features [34, 10]. Furthermore, the majority of studies focus on the field of Language learning [3, 9] which is inherently different to factual knowledge or reading comprehension 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 [36]. 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 USMLE dataset) [7].

Most similar to our work is the study by Loginova et al. [12], 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. Whereas Loginova et al. focus on training and calibrating Encoder-Only Question-Answering models, we instead examine "off-the-shelf" Decoder-Only models, which inherently incorporate a greater amount of factual knowledge as a byproduct of their language modeling objective [37]. Additionally, we broaden the research scope beyond language comprehension to include factual knowledge understanding, through the use of three datasets that assess knowledge across different education levels. Finally, rather than relying on proxied comparative assessments – where assessors classify questions as "easy" or "difficult" – we leverage fine-grained, continuous difficulty labels, such as the proportion of students answering a question correctly.

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 [39, 16]. For the task of question difficulty estimation, this paradigm has been explored in the context of language comprehension by with some success [23]. More recently Raina et al. achieved good results in the USMLE and CMCQRD datasets through comparative assessments (i.e., given two question items, the LLM's task is to determine which is more difficult) [24].

The present work is also influenced by the work by Zotos et al., where a variety of analyses showed a weak, but statistically significant, correlation between human and machine perceived difficulty [40]. 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 three distinct question sets assessing both factual knowledge and reading comprehension.

3. DATA

The three MCQ datasets that we use in our experiments are described more in detail in the following subsections. The first is a dataset on the domain of Biopsychology that is not publicly available. The second is the publicly available dataset used in the BEA 2024 Shared Task [36]. Lastly, we also use the "Cambridge Multiple-Choice Questions Reading Dataset" [15]. For brevity, we refer to the "Biopsychology", "USMLE" and "CMCQRD" datasets respectively. Our choice is driven by the requirement of having questionsets along with students selection rates (serving as proxies for item difficulty scores). Considering that, to the best of our knowledge, the USMLE and CMCQRD datasets are the only publicly available resources satisfying this requirement. Additionally, we use a non-publicly available dataset as a complement. This choice is in line with the observation by AlKhuzaey et al., who note that most studies tackling this task resort to using private datasets [1].

As will be explained in Sections 3.1 through 3.3, the three datasets vary in multiple aspects, for example question formulation, number of incorrect choices (also known as 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 and CMCQRD datasets, as shown in Table 2.

[1] Code available at: `https://github.com/LeonidasZotos/Are-You-Doubtful-Oh-It-Might-Be-Difficult-Then`.

Table 1: Examples questions from the Biopsychology, USMLE and CMCQRD datasets. Correct answer in green. Two examples are given from the Biopsychology dataset to illustrate the phrasing variability.

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
USMLE	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
CMCQRD	Mum caught Jess by the arm. 'Come with me,' she said. Jess followed her through to the study and there on the table, propped against the wall, was an unframed painting, unmistakably one of Grandpa's, yet unlike anything he had done before, and clearly nowhere near finished. 'Do you know anything about this?' said Mum. Jess shook her head. 'I've never seen it before. I didn't know he was working on anything.' [...] And it was hard to imagine anyone, even a goddess, having any influence over someone as wilful as Grandpa. 'He doesn't need me to inspire him,' she said. 'He's been painting all his life.' What impression is given of Jess's grandfather in the final paragraph?	ⓐ He lacks confidence in his ability. ⓑ He values Jess's opinions. ⓒ He has a strong character. ⓓ He finds it hard to concentrate.

Table 2: Train and Test splits as used in our experiments. For USMLE, we use the splits as provided in the competition [36]. For the Biopsychology and CMCQRD datasets, we randomly sampled the questions, keeping the same percentage of training/testing samples as in USMLE. For the CMCQRD, we also ensured that text passages present in the train set were not present in the test set.

Dataset	Train	Test	Total	Public?
Biopsychology	573	246	819	✘
USMLE	466	201	667	✔
CMCQRD	552	241	793	✔

The item difficulty labels differ between the three datasets. In the Biopsychology and CMCQRD datasets, difficulty is measured by the proportion of students who answered correctly (a higher value indicates an easier question). This measure is commonly known as the *p-value* of a question item [30]. In contrast, the USMLE 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 USMLE 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 USMLE difficulty scores to their complements such that they also reflect the proportion of correct responses per question.

Finally, CMCQRD is also equipped with an Item Response Theory (IRT)-based metric, specifically scaled single parameter Rasch Model difficulty estimates. To the best of our knowledge, while previous results exist using this IRT-based metric [24], no prior work has attempted to estimate p-values for this dataset. In this work, we conduct experiments independently estimating both the p-value and the IRT-based metric, achieving state-of-the-art results with the latter, where prior results are also available.

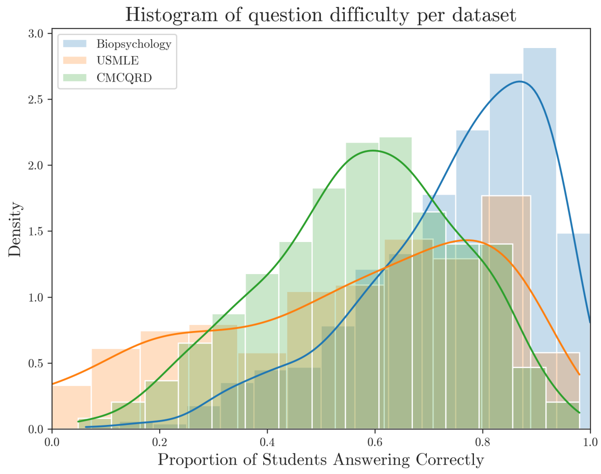


Figure 2: Distribution of question difficulty, based on the proportion of students answering each question correctly. For the USMLE dataset, only transformed p-values are available.

Figure 2 illustrates the distribution of difficulty across the three datasets (focusing on the p-value for the CMCQRD and not the IRT metric). As shown, while the datasets contain questions of varying difficulty levels, they are generally skewed toward easier questions. This trend is particularly evident in the Biopsychology dataset, where 81% of questions were answered correctly by at least 60% of the student population.

3.1 Biopsychology

The Biopsychology dataset originates from a course taught in the 1st year of the Psychology undergraduate degree at the Social Sciences Faculty of the University of Groningen, covering content from the classic textbook “Biological Psychology” by Kalat [11]. 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 1. Given that LLMs demonstrate sensitivity to input formulation [4], the presence of such variability in the data improves generalisation of our method across datasets.

3.2 USMLE

The United States Medical Licensing Examination (USMLE) question set was developed by the National Board of Medical Examiners (NBME) and Federation of State Medical Boards (FSMB) [36]. 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 1. As can be seen, questions of this dataset are longer (755 characters compared to 103 characters for the Biopsychology set) and are of technical nature.

3.3 CMCQRD

The Cambridge Multiple-Choice Questions Reading Dataset (CMCQRD) consists of 120 text passages, each accompanied with a number of MCQs (total of 793 question items) aiming at evaluating the student’s language comprehension abilities. While the questions target various Common European Framework of Reference for Languages (CEFR) proficiency levels (B1 to C2), we do not leverage this additional information in our experiments. Furthermore, compared to the Biopsychology and USMLE datasets, question items in the CMCQRD are the longest, with an average of $3,618$ characters per item.

4. APPROACH

Given an MCQ, the task is to predict its difficulty measured by the proportion of students that select the correct choice (in addition to the IRT-based metric available for the CMCQRD dataset). An 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[2] which receives as input a vectorised representation of the MCQ, as well the uncertainty of multiple LLMs answering the same MCQ[3]. We opted for a relatively simple Random Forest Regressor, as it allows for analysis using explainability methods (through SHAP post-hoc analysis), while still effectively demonstrating the usefulness of model uncertainty in this context. As features, we use Textual Features and Model Uncertainty, as described in sections 4.1 and 4.2. We further supplement this basic system using higher-level textual features namely Linguistic Features and Choice Similarity, described in section 4.3.

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 [27]. 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,

[2] As provided by the Scikit-Learn Library, using the default hyper-parameters [18].

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

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 [5] and domain-specific Bioclinical BERT Embeddings [2]. The Bioclinical BERT Embeddings, previously also employed by team ITEC in the BEA 2024 shared task [29], 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. In our experiments, Bioclinical BERT embeddings consistently yielded poorer results, so we omit them here for brevity.

4.2 Model Uncertainty

The current 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 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 any additionally textual, linguistic or choice similarity features are (optionally) also added. This vector is then fed to the regressor.

1st Token Probability. The first method to measure model uncertainty is by inspecting the softmax probability of the 1st token to be generated as the answer 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 [33, 31, 32, 38], we create ten random different orderings for each question and let the model answer each MCQ ten times[4]. 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 4.4 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 to 1.

Choice-Order Sensitivity. Pezeshkpour and Hruschka [21] 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 the correct answer. 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.

4.3 Additional Features

We also conduct experiments with two additional feature sets: linguistic and choice similarity features. First, based on the work of Ha et al. [8], we extract 17 higher-level linguistic features from each question item. These features vary in complexity, ranging from the number of sentences to the occurrence of additive connectives. Moreover, we also develop a simple baseline that uses these linguistic features along with `word2vec` embeddings [14], similar to the work by Yaneva et al. [35][5].

In the current work we also explore the use of choice similarity defined (per MCQ) as the average cosine similarity between each distractor and the correct answer choice, similar to the approach by Susanti et al. [28]. This is operationalised using the Sentence Transformer library [25] with one of two models: `all-MiniLM-L6-v2`[6] (general embeddings) and `S-PubMedBert-MS-MARCO`[7] (medical/health text domain embeddings). The two setups are henceforth referred to as "General Similarity" and "Medical Similarity" respectively.

4.4 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 modeling objective [37], 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 use the models' instruction-tuned (also known as "chat") variant and experiment with both default precision and 4-bit quantised models. This comparison is important because, although quantized models are more efficient (albeit slightly less capable), it is unclear whether they are also more miscalibrated[8]. To adapt them for the task of MCQ answering, we use the instruction prompt in Figure 3 based on the experimental setup of Plaut et al. [22] as well as Zotos et al. [40].

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

[5] It is worth noting that in contrast to Yaneva et al. [35], we do not use readability metrics as some question items do not meet the 100 word requirement for these metrics to be computed.

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

[7] `https://huggingface.co/pritamdeka/S-PubMedBert-MS-MARCO`.

[8] Details of the models used are in Table 3.

Table 3: Default precision and quantised LLMs used in the experiments. All models can be found at the HuggingFace Hub.

Model Name	Default Precision	4-bit Quantisation
phi3_5-chat	`microsoft/Phi-3.5-mini-instruct`	`unsloth/Phi-3.5-mini-instruct-bnb-4bit`
Llama3_2-3b-chat	`meta-llama/Llama-3.2-3B-Instruct`	`unsloth/Llama-3.2-3B-Instruct-bnb-4bit`
Qwen2_5-3b-chat	`Qwen/Qwen2.5-3B-Instruct`	`unsloth/Qwen2.5-3B-Instruct-bnb-4bit`
Llama3_1-8b-chat	`meta-llama/Llama-3.1-8B-Instruct`	`unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit`
Qwen2_5-14b-chat	`Qwen/Qwen2.5-14B-Instruct`	`unsloth/Qwen2.5-14B-Instruct-bnb-4bit`
Qwen2_5-32b-chat	`Qwen/Qwen2.5-32B-Instruct`	`unsloth/Qwen2.5-32B-Instruct-bnb-4bit`
Yi-34b-chat	`01-ai/Yi-34B-Chat`	`unsloth/yi-34b-chat-bnb-4bit`
Llama3_1-70b-chat	`meta-llama/Llama-3.1-70B-Instruct`	`unsloth/Meta-Llama-3.1-70B-Instruct-bnb-4bit`
Qwen2_5-72b-chat	`Qwen/Qwen2.5-72B-Instruct`	`unsloth/Qwen2.5-72B-Instruct-bnb-4bit`

Table 4: 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. Best overall results are highlighted in light green. All results are averaged over ten repetitions, with the standard deviation not exceeding 0.002.

Method	Biopsychology		USMLE		CMCQRD		CMCQRD_IRT	
Baselines								
Dummy Regressor	0.1667		0.3110		0.1833		9.7568	
Best Literature Result [24]	-		0.291		-		8.5	
Only Text								
TF-IDF	0.1479		0.3092		0.1843		9.2531	
BERT Embeddings	0.1498		0.3066		0.1843		8.9347	
Only Uncertainty								
1st Token Probabilities	0.1473		0.3041		0.1770		9.4161	
Choice-Order Sensitivity	0.1543		0.3155		0.1788		9.9787	
Both Uncertainty Features	0.1460		0.3034		0.1754		9.3705	
Text and Uncertainty								
	TF-IDF	**BERT**	**TF-IDF**	**BERT**	**TF-IDF**	**BERT**	**TF-IDF**	**BERT**
First Token Probability	0.1361	0.1362	0.3037	0.2868	0.1668	0.1669	8.5661	**8.2763**
Choice-Order Sensitivity	0.1378	0.1409	0.310	0.2906	0.1658	0.1676	8.6663	8.4204
Both Uncertainty Features	**0.1359**	0.1365	0.3044	**0.2864**	**0.1654**	0.1669	8.5933	8.3009

> **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 3: 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*.

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 text and model uncertainty feature sets (Section 5.1), followed by an exploration of the effect of additional features (Section 5.2). We conclude with a post-hoc analysis of our system using SHAP explanations (Section 5.3). While the following sections present only the relevant results, Appendix B provides an overview of the results for all experiments. All experiments were conducted using two Nvidia A100 GPUs.

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 [18], as used in the BEA 2024 Shared task. As previously mentioned, we use a Random Forest 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 three datasets in Table 4. Similarly, Table 4 presents the performance of the default precision variants of the models, while the results for the quantised models are reported in Appendix B. Lastly, for CMCQRD, we report the performance on pre-

Table 5: Root Mean Squared Error (RMSE, the lower the better) on the test set using additional features. Lowest achieved RMSE per dataset is shown in boldface. Best overall results, also in comparison to those presented in Table 4, are additionally highlighted in light green. All results are averaged over ten repetitions, with the standard deviation not exceeding 0.002.

Method	Biopsychology		USMLE		CMCQRD		CMCQRD_IRT	
Baselines								
Dummy Regressor	0.1667		0.3110		0.1833		9.7568	
Best Literature Result [24]	-		0.291		-		8.5	
Linguistic Features Baseline [8]	0.1544		0.3147		0.1852		9.3335	
Only Choice Similarity								
General (`all-MiniLM-L6-v2`)	0.1895		0.3567		0.2226		12.2829	
Medical (`S-PubMedBert-MS-MARCO`)	0.1883		0.3432		0.2146		11.4768	
	TF-IDF	**BERT**	**TF-IDF**	**BERT**	**TF-IDF**	**BERT**	**TF-IDF**	**BERT**
Text, Both Uncertainties & Choice Similarity								
General Similarity	0.1372	0.1367	0.2835	0.2862	**0.1651**	0.1669	8.5934	8.2956
Medical Similarity	0.1376	**0.1361**	0.2836	0.2853	**0.1651**	0.1664	8.5999	**8.2325**
Both Similarities	0.1378	0.1365	0.2847	0.2862	0.1653	0.1676	8.5694	8.2817
Text, Both Uncertainties, Both Sim & Linguistic Features								
	0.1393	0.1362	**0.2817**	0.2857	0.1652	0.1673	8.5490	8.6575

dicting both p-values and the IRT-based metric.

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, and the achieved RMSE scores of the top 10 approaches were within 0.009. Even so, there are consistent differences between the experimental setups. Most importantly for this research, incorporating model uncertainty alongside text features significantly reduces RMSE across all datasets, outperforming the best scores in previous literature. Even lower RMSE is achieved for the USMLE and CMCQRD datasets when additional features are included (see Section 5.2). Furthermore, our exploration revealed that 1st Token Probability consistently serves as a more useful signal for the task than Choice-Order Sensitivity. However, combining both yields the best results. Regarding the two text vectorization methods, we find no significant differences between them, except for the USMLE dataset, where BERT Embeddings outperform TF-IDF scores.

5.2 Effect of Additional Features

Naturally, additional features of an MCQ might capture aspects that make the question easier or more challenging. Table 5 explores the effect of higher-level linguistic features and the similarity between choices (as described in Section 4.3). Expectedly, using choice similarity as the only feature yields poor results on the task. Similarly, the baseline that focuses on linguistic features only marginally beats the Mean Regressor baseline for two of the four experiments. However, combining either, or both, of these features with text and model uncertainty features further improves our best results for all but the Biopsychology experiments (where the best result is still achieved without the additional features). Moreover, we observe that there is no clear advantage between the General or Medical Similarity, with the latter also seemingly being useful in the non-medical domain (CMCQRD dataset). These experiments highlight that while model uncertainty and basic text encodings (such as BERT or TF-IDF scores) capture certain aspects of difficulty, other factors should still be utilized for this task.

5.3 Feature Importance

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 [13]. To maintain conciseness, we present SHAP summary plots for a selected subset of experiments that we found to be the most insightful.

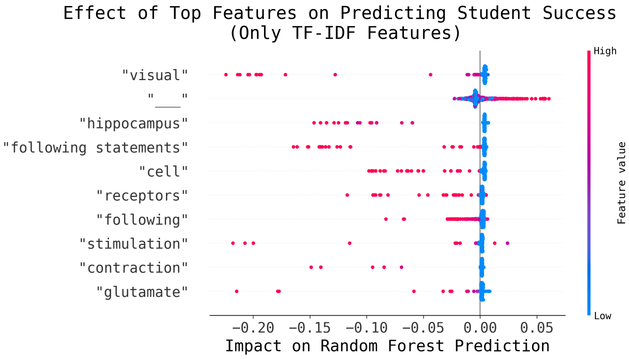
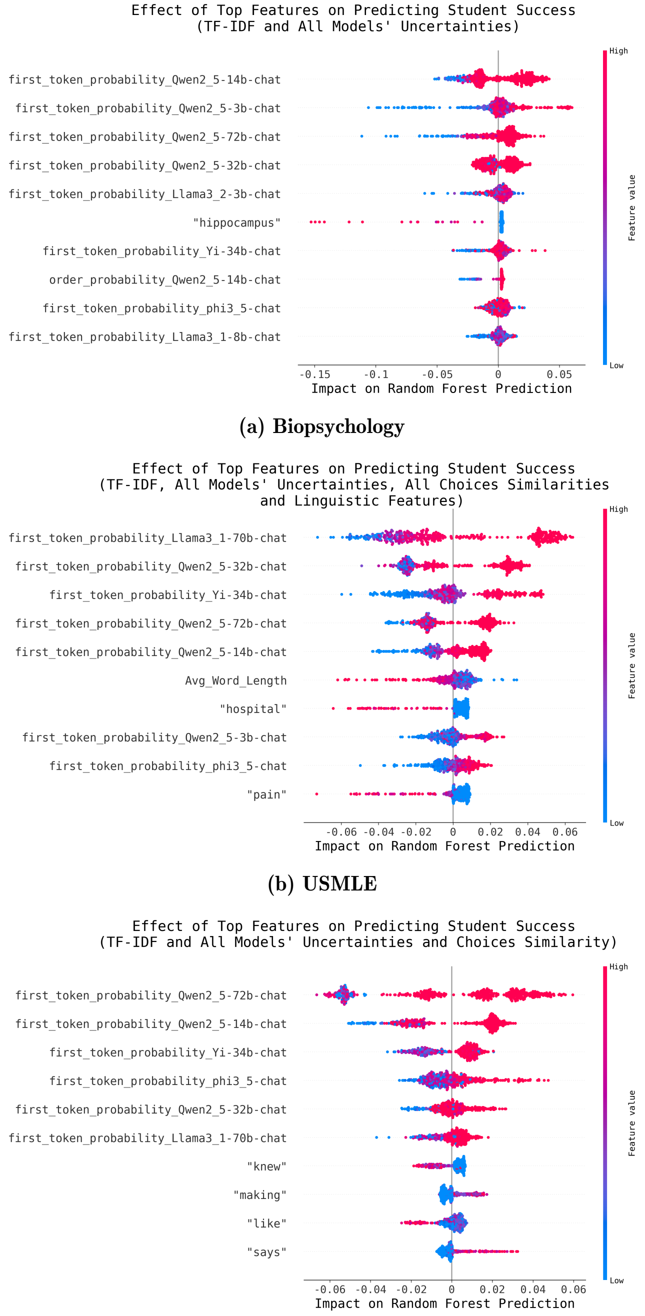


Figure 4: Biopsychology Dataset. Shapley summary plot showing the contribution of the top ten uni/bi-gram features to the Random Forest's predictions, highlighting their importance and impact direction. Features are ranked by their average influence, with dots representing individual question items and colour indicating TF-IDF scores. Results averaged over ten repetitions.

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 is useful because it allows us to gain an overview of the influence of lexical features before introduc-



(a) Biopsychology

(b) USMLE

(c) CMCQRD (p-value)

Figure 5: Shapley summary plots for the three 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.

ing model uncertainty, while also highlighting any features that unexpectedly influence question difficulty (e.g., interrogative words).

This analysis relies on TF-IDF scores, as BERT Embeddings cannot directly be traced back to individual words. Figure 4 shows the ten most impactful features, along with their effect on the Regressors' prediction for each MCQ item. High TF-IDF scores are highlighted in red and broadly represent the presence of the word in an item. Furthermore, the impact on the Regressor's prediction is either positive or negative, meaning that a feature can either lower or increase the predicted difficulty score. As can be seen, the presence of certain terms (e.g., "visual", "hippocampus") lead the Regressor to predict higher 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 shows that the presence of certain words can steer the Regressor towards predicting a higher or lower difficulty, we are mostly interested in the contribution of features related to model uncertainty. Figures 5a, 5b and 5c 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, USMLE and CMCQRD datasets, respectively. It is worth noting that for the CMCQRD dataset, we focus on the prediction of the number of students selecting the correct rate (p-value), instead of the IRT-based metric.

In all 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 each configuration, the uncertainty of different models has the greatest influence. Notably, the uncertainty of Qwen models consistently serves as a strong indicator of difficulty. Furthermore, this analysis hints towards model size being important, especially when comparing the Biopsychology and the USMLE results: For the latter, the confidence of larger models is more influential in the Regressor's prediction of question difficulty. 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 leveraged for the task of question item difficulty estimation using three MCQ datasets focusing on factual knowledge and language comprehension. We demonstrate, in experimental setups of varying complexity, 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. Additionally, our results suggest that an LLM's uncertainty does not fully account for certain aspects of item difficulty, such as linguistic complexity[9].

Our methodological design is intentionally simple, serving as a proof of concept for this approach. This simplicity stems from various design choices. Firstly, 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. Secondly, we conducted a series of experiments using features of varying complexity to demonstrate that, while incorporating additional features (e.g., average cosine similarity between each distractor and the correct answer choice) can improve task performance, the improvements are often marginal compared to relying solely on a simple text encoding combined with model uncertainty.

Shifting into a broader perspective, our findings suggest that there are similarities between the way LLMs and students process educational material. While caution is necessary, we see potential for future research to leverage LLMs for student and cohort modeling.

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. Naturally, this 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 [4, 26].

Lastly, due to dataset availability, we evaluated our approach solely on three examinations. It remains unclear whether model uncertainty could also help assess the difficulty of exams in other skill sets, such as mathematical reasoning.

References

[1] S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914, Sep 2024.

[2] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. In A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[3] S. Bi, X. Cheng, Y.-F. Li, L. Qu, S. Shen, G. Qi, L. Pan, and Y. Jiang. Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4645–4654, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[4] S. Biderman, H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, P. S. Ammanamanchi, S. Black, J. Clive, A. DiPofi, J. Etxaniz, B. Fattori, J. Z. Forde, C. Foster, J. Hsu, M. Jaiswal, W. Y. Lee, H. Li, C. Lovering, N. Muennighoff, E. Pavlick, J. Phang, A. Skowron, S. Tan, X. Tang, K. A. Wang, G. I. Winata, F. Yvon, and A. Zou. Lessons from the Trenches on Reproducible Evaluation of Language Models, May 2024.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *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, June 2019. Association for Computational Linguistics.

[6] M. Gierl, O. Bulut, Q. Guo, and X. Zhang. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87:0034654317726529, 08 2017.

[7] S. Gombert, L. Menzel, D. Di Mitri, and 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, and Z. Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 483–492, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[8] L. A. Ha, V. Yaneva, P. Baldwin, and J. Mee. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, and T. Zesch, editors, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[9] Notably, a question item can be complex yet easy, or vice versa [19].

[9] J. He, L. Peng, B. Sun, L. Yu, and Y. Zhang. Automatically predict question difficulty for reading comprehension exercises. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1398–1402. IEEE, 2021.

[10] F.-Y. Hsu, H.-M. Lee, T.-H. Chang, and Y.-T. Sung. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing and Management*, 54:969–984, 11 2018.

[11] J. W. Kalat. *Biological psychology*. Cengage Learning, 2016.

[12] E. Loginova, L. Benedetto, D. Benoit, and P. Cremonesi. Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. In R. Mitkov and G. Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 846–855, Held Online, Sept. 2021. INCOMA Ltd.

[13] S. M. Lundberg and S.-I. Lee. 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., 2017.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[15] A. Mullooly, O. Andersen, L. Benedetto, P. Buttery, A. Caines, M. J. F. Gales, Y. Karatay, K. Knill, A. Liusie, V. Raina, and S. Taslimipoor. *The Cambridge Multiple-Choice Questions Reading Dataset*. Cambridge University Press and Assessment, 2023.

[16] Q. Pan, Z. Ashktorab, M. Desmond, M. S. Cooper, J. Johnson, R. Nair, E. Daly, and W. Geyer. Human-centered design recommendations for llm-as-a-judge, 2024.

[17] J. Papoušek, V. Stanislav, and R. Pelánek. 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, 2016.

[18] 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] R. Pelánek, T. Effenberger, and J. Čechák. Complexity and difficulty of items in learning systems. *International Journal of Artificial Intelligence in Education*, 32:1–37, 05 2021.

[20] K. Perkins, L. Gupta, and R. Tammana. Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language Testing*, 12(1):34–53, 1995.

[21] P. Pezeshkpour and E. Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023.

[22] B. Plaut, K. Nguyen, and T. Trinh. Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a, 2024.

[23] V. Raina and M. Gales. Question Difficulty Ranking for Multiple-Choice Reading Comprehension, Apr. 2024.

[24] V. Raina, A. Liusie, and M. Gales. Finetuning LLMs for comparative assessment tasks. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3345–3352, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics.

[25] N. Reimers and I. Gurevych. 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, 11 2019.

[26] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2023.

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

[28] Y. Susanti, T. Tokunaga, and H. Nishikawa. Integrating automatic question generation with computerised adaptive test. *Research and Practice in Technology Enhanced Learning*, 15:1–22, 2020.

[29] A. Tack, S. Buseyne, C. Chen, R. D'hondt, M. De Vrindt, A. Gharahighehi, S. Metwaly, F. K. Nakano, and A.-S. Noreillie. ITEC at BEA 2024 shared task: Predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 512–521, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[30] G. van de Watering and 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(2):133–147, 2006.

[31] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators, 2023.

[32] X. Wang, C. Hu, B. Ma, P. Röttger, and B. Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think, 2024.

[33] S.-L. Wei, C.-K. Wu, H.-H. Huang, and H.-H. Chen. Unveiling selection biases: Exploring order and token sensitivity in large language models, 2024.

[34] K. Xue, V. Yaneva, C. Runyon, and P. Baldwin. Predicting the difficulty and response time of multiple choice questions using transfer learning. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, editors, *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics.

[35] V. Yaneva, D. Jurich, L. A. Ha, and P. Baldwin. Using linguistic features to predict the response process complexity associated with answering clinical MCQs. In J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, editors, *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 223–232, Online, Apr. 2021. Association for Computational Linguistics.

[36] V. Yaneva, K. North, P. Baldwin, L. A. Ha, S. Rezayi, Y. Zhou, S. Ray Choudhury, P. Harik, and B. Clauser. Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Questions. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, and Z. Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[37] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. A survey of large language models, 2023.

[38] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large language models are not robust multiple choice selectors, 2024.

[39] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[40] L. Zotos, H. van Rijn, and 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, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11304–11316, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics.

APPENDIX

A. MODEL CORRECTNESS AND UNCERTAINTY

Table 6 presents the performance of each model on the three question sets, as well as the relation between model certainty and model correctness. In line with the results of Plaut et al. [22], it is clear that both tested metrics correlate well with model correctness: On average, the mean certainty for the chosen option is higher for the correctly answered question items. This suggests that the two metrics indeed capture an aspect of model certainty. Lastly, we can see that this trend persists despite quantisation, although quantised models generally exhibit lower performance compared to their default precision counterparts.

Table 6: Model correctness and answer probability in terms of Mean 1st Token and Choice-Order Probability in the Biopsychology, USMLE and CMCQRD question sets. "Overall Correctness" indicates the proportion of correctly answered questions, while the probabilities in blue and red indicate the mean model certainty for the model's choice for correctly and incorrectly answered questions respectively. As can be seen, on average, model certainty is higher when questions are answered correctly, especially for larger LLMs.

		Default Precision			4-bit Quantisation		
Dataset	Model	Correctness	Mean Probability		Correctness	Mean Probability	
			1st Token	Choice-Order		1st Token	Choice-Order
Biopsychology	phi3_5	0.821	0.921 / 0.696	0.940 / 0.743	0.288	0.418 / 0.397	0.494 / 0.455
	Llama3_2-3b	0.740	0.747 / 0.541	0.849 / 0.676	0.714	0.735 / 0.509	0.855 / 0.652
	Qwen2_5-3b	0.772	0.830 / 0.609	0.856 / 0.641	0.797	0.857 / 0.610	0.874 / 0.642
	Llama3_1-8b	0.817	0.641 / 0.445	0.863 / 0.648	0.832	0.677 / 0.455	0.847 / 0.596
	Qwen2_5-14b	0.896	0.968 / 0.790	0.972 / 0.823	0.897	0.971 / 0.777	0.973 / 0.800
	Qwen2_5-32b	0.934	0.972 / 0.759	0.981 / 0.802	0.933	0.968 / 0.750	0.978 / 0.792
	Yi-34b	0.880	0.888 / 0.629	0.917 / 0.681	0.874	0.868 / 0.582	0.896 / 0.615
	Llama3_1-70b	0.933	0.938 / 0.649	0.975 / 0.794	0.935	0.945 / 0.690	0.977 / 0.811
	Qwen2_5-72b	0.941	0.971 / 0.808	0.981 / 0.875	0.939	0.962 / 0.750	0.982 / 0.815
USMLE	phi3_5	0.571	0.781 / 0.612	0.838 / 0.703	0.189	0.319 / 0.306	0.363 / 0.372
	Llama3_2-3b	0.634	0.618 / 0.428	0.785 / 0.614	0.658	0.596 / 0.407	0.767 / 0.572
	Qwen2_5-3b	0.477	0.670 / 0.563	0.740 / 0.648	0.514	0.663 / 0.529	0.740 / 0.630
	Llama3_1-8b	0.627	0.371 / 0.290	0.637 / 0.507	0.679	0.431 / 0.325	0.730 / 0.585
	Qwen2_5-14b	0.744	0.897 / 0.699	0.911 / 0.750	0.732	0.898 / 0.690	0.906 / 0.739
	Qwen2_5-32b	0.811	0.898 / 0.659	0.923 / 0.744	0.816	0.906 / 0.661	0.931 / 0.744
	Yi-34b	0.652	0.777 / 0.575	0.831 / 0.666	0.652	0.726 / 0.513	0.781 / 0.572
	Llama3_1-70b	0.885	0.849 / 0.493	0.946 / 0.679	0.886	0.841 / 0.485	0.941 / 0.666
	Qwen2_5-72b	0.849	0.930 / 0.668	0.948 / 0.726	0.858	0.909 / 0.645	0.939 / 0.729
CMCQRD	phi3_5	0.706	0.818 / 0.664	0.845 / 0.712	0.246	0.374 / 0.369	0.397 / 0.417
	Llama3_2-3b	0.666	0.671 / 0.493	0.806 / 0.635	0.705	0.688 / 0.518	0.829 / 0.679
	Qwen2_5-3b	0.662	0.791 / 0.622	0.816 / 0.659	0.744	0.864 / 0.673	0.884 / 0.704
	Llama3_1-8b	0.752	0.583 / 0.415	0.800 / 0.602	0.781	0.650 / 0.454	0.828 / 0.628
	Qwen2_5-14b	0.888	0.950 / 0.737	0.958 / 0.747	0.894	0.951 / 0.740	0.957 / 0.770
	Qwen2_5-32b	0.908	0.954 / 0.728	0.970 / 0.785	0.912	0.949 / 0.743	0.969 / 0.813
	Yi-34b	0.841	0.873 / 0.624	0.900 / 0.668	0.831	0.794 / 0.546	0.831 / 0.582
	Llama3_1-70b	0.917	0.926 / 0.684	0.973 / 0.809	0.912	0.912 / 0.617	0.972 / 0.789
	Qwen2_5-72b	0.914	0.962 / 0.737	0.971 / 0.766	0.912	0.955 / 0.739	0.973 / 0.793

B. RESULTS OVERVIEW

Table 7 presents an overview of the results of all experiments.

Table 7: Root Mean Squared Error (RMSE, the lower the better) on the test set using different sets of features. Best overall results are highlighted in light green. All results are averaged over ten repetitions, with the standard deviation not exceeding 0.002.

Method	Biopsychology		USMLE		CMCQRD		CMCQRD_IRT	
Baselines								
Dummy Regressor	0.1667		0.3110		0.1833		9.7568	
Best Literature Result [24]	-		0.291		-		8.5	
Linguistic Features Baseline [8]	0.1544		0.3147		0.1852		9.3335	
Only Text								
TF-IDF	0.1479		0.3092		0.1843		9.2531	
BERT Embeddings	0.1498		0.3066		0.1843		8.9347	
Only Uncertainty								
	Default	**4-bit**	**Default**	**4-bit**	**Default**	**4-bit**	**Default**	**4-bit**
1st Token Probabilities	0.1473	0.1539	0.3041	0.2960	0.1770	0.1752	9.4161	9.5385
Choice-Order Sensitivity	0.1543	0.1582	0.3155	0.3178	0.1788	0.1880	9.9787	9.8563
Both Uncertainty Features	0.1460	0.1538	0.3034	0.2968	0.1754	0.1761	9.3705	9.4899
Only Choice Similarity								
General (`all-MiniLM-L6-v2`)	0.1895		0.3567		0.2226		12.2829	
Medical (`S-PubMedBert-MS-MARCO`)	0.1883		0.3432		0.2146		11.4768	
	TF-IDF	**BERT**	**TF-IDF**	**BERT**	**TF-IDF**	**BERT**	**TF-IDF**	**BERT**
Text and Uncertainty								
First Token Probability (Default)	0.1361	0.1362	0.3037	0.2868	0.1668	0.1669	8.5661	8.2763
Choice-Order Sensitivity (Default)	0.1378	0.1409	0.3100	0.2906	0.1658	0.1676	8.6663	8.4204
Both Uncertainty Features (Default)	0.1359	0.1365	0.3044	0.2864	0.1654	0.1669	8.5933	8.3009
First Token Probability (4-bit)	0.1365	0.1385	0.2851	0.2854	0.1680	0.1675	8.5392	8.3322
Choice-Order Sensitivity (4-bit)	0.1309	0.1411	0.2951	0.2961	0.1671	0.1672	8.6218	8.3459
Both Uncertainty Features (4-bit)	0.1371	0.1388	0.2856	0.2846	0.1685	0.1670	8.5587	8.3728
Text, Both Uncertainties & Choice Similarity								
General Sim. (Default)	0.1372	0.1367	0.2835	0.2862	0.1651	0.1669	8.5934	8.2956
Medical Sim. (Default)	0.1376	0.1361	0.2836	0.2853	0.1651	0.1664	8.5999	8.2325
Both Sim. (Default)	0.1378	0.1365	0.2847	0.2862	0.1653	0.1676	8.5694	8.2817
General Sim. (4-bit)	0.1386	0.1389	0.2850	0.2841	0.1674	0.1659	8.5696	8.3338
Medical Sim. (4-bit)	0.1378	0.1381	0.2856	0.2844	0.1676	0.1660	8.5204	8.2801
Both Sim. (4-bit)	0.1397	0.1412	0.2860	0.2850	0.1673	0.1659	8.5475	8.2820
Text, Both Uncertainties, Both Sim & Linguistic Features								
Default	0.1393	0.1362	0.2817	0.2857	0.1652	0.1673	8.5490	8.6575
4-bit	0.1411	0.1410	0.2853	0.2844	0.1677	0.1660	8.4966	8.5907