BERT-Based Classification of Cross-Linguistic MCQs According to the Bloom Taxonomy

João Francisco Botas[1]
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

Ana Rita Peixoto
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

Eugénio Ribeiro
INESC-ID Lisboa, Portugal
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

Abstract

The Bloom Taxonomy provides a useful framework for describing the cognitive demands of Multiple Choice Questions (MCQs), yet automatically assigning Bloom levels remains challenging due to category overlap and the limited evidence of higher-order thinking in MCQs. This study aims to evaluate transformer-based models, including BERT and ModernBERT variants, to enhance their performance on a four-level Bloom classification task across multiple scenarios, while also conducting model interpretability and error analysis.

Across 40 settings tested, BERT models consistently outperform a keyword-based baseline, highlighting the contextual and semantic representations for this task. Performance is broadly similar between BERT and ModernBERT, with only minor differences across architectures, while the inclusion of answer options yields only small gains, suggesting that the question stem alone contains most of the discriminative signal. Besides, cross-linguistic experiments show comparable results between the original English data and Portuguese, although translated data exhibits a slight performance degradation, likely due to direct translation noise and subtle syntactic shifts. Errors are concentrated between adjacent Bloom levels, and models perform best on the *Remembering* level. In contrast, the *Applying* level remains the most difficult class to predict, largely because of class imbalance and conceptual overlap.

The findings indicate that our approach constitutes a stable pipeline for Bloom classification, but its performance remains constrained by the intrinsic ambiguity of the taxonomy and the structural characteristics of the available data.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases Bloom Taxonomy, Multiple Choice Questions, Transformers, Multi-Language, Natural Language Processing, Educational Assessment

Digital Object Identifier 10.4230/OASIcs.SLATE.2026.4

Funding This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT), ISTAR-Iscte Pluriannual Projects UID/04466/2025 (`https://doi.org/10.54499/UID/04466/2025`) and INESC-ID Projects UID/50021/2025 (`https://doi.org/10.54499/UID/50021/2025`) and UID/PRR/50021/2025 (`https://doi.org/10.54499/UID/PRR/50021/2025`).

1 Introduction

Assessing the difficulty and complexity of educational questions remains a challenging task, particularly in large-scale settings where Multiple Choice Question (MCQ) are widely used. Traditional approaches rely on pretesting procedures, pilot studies, or expert judgment

[1] Corresponding author

to assign difficulty levels, but these methods are costly, time-consuming, and difficult to scale [11]. As a result, there is a growing interest in automated approaches that can infer question characteristics directly from textual and structural features, enabling more efficient analysis of large educational datasets.

A key dimension in understanding question complexity is not only in "how difficult a question is", but also "what type of cognitive process it requires to be answered". In this context, Bloom's Taxonomy (BT) provides a structured framework to characterize the level of thinking involved in answering a question, organizing cognitive skills into six hierarchical levels: 1-*Remembering*, 2-*Understanding*, 3-*Applying*, 4-*Analyzing*, 5-*Evaluating*, and 6-*Creating* (for the Revised Bloom's Taxonomy (BT)) [3]. Unlike traditional difficulty measures, which are often derived from exhaustive student performance data, BT focuses on the item's intended cognitive demand, making it suitable for analyzing question design [24]. However, applying this framework in practice is non-trivial due to the cognitive levels often overlapping, being context-dependent, and being subjective to interpret, including for specific types of questions, such as MCQs [22]. Despite these challenges, BT offers a meaningful proxy for complexity that can be operationalized through multiple Natural Language Processing (NLP) techniques, from the more traditional to the most recent, enabling automated analysis of educational content [18].

In this study, we investigate the use of transformer-based models, specifically BERT architectures [6], to automatically classify MCQs according to BT levels. Our approach focuses on evaluating how well these models capture the semantic and contextual information required to distinguish between cognitive levels, using both the question text and, optionally, the answer options as input. Furthermore, we extend this analysis to a multilingual setting by comparing model performance across English and Portuguese versions of the same dataset, using specific BERT models for each. This allows us to examine the impact of language variation on model behavior, particularly when data is obtained through automatic translation. In addition to standard classification metrics, we also consider evaluation strategies that account for the ordinal nature of Bloom levels, providing a more nuanced understanding of model performance. We also try to analyze how different question formats (fill-in-blank versus non-fill-in-blank MCQs) and domain can affect classification performance and model global behavior. All of these aspects are systematically evaluated across multiple experimental configurations, enabling a comprehensive comparison of modeling choices and their impact on performance.

To guide this study, we define a set of Research Questions (RQs) organized around three main topics: model effectiveness, interlingual robustness, and error analysis. These research questions are designed to assess not only the overall performance of the proposed approaches but also to better understand their behavior under different configurations, languages, and data characteristics. The defined RQs are as follows:

1. To what extent can BERT-based models accurately classify MCQs according to Bloom's Taxonomy (BT), and how do different configurations affect performance?
2. Do models trained on translated datasets maintain comparable performance across languages, and how do linguistic and syntactic variations influence classification outcomes?
3. What are the most common misclassification patterns, particularly between adjacent cognitive levels, and how are these affected by other data typologies?

The remainder of this document is organized as follows: Section 2 reviews related work on Bloom's Taxonomy (BT) classification, MCQ assessment, and NLP approaches. Section 3 describes the datasets, preprocessing steps, model configurations, experimental

setup, and evaluation metrics. Section 4 presents the results, starting with the baseline and then analyzing the transformer-based models, including cross-linguistic comparisons, error patterns, and profiling analyses. Finally, Section 5 summarizes the main findings, limitations, and future work.

2 Related Work

Bloom's Taxonomy (BT) has been widely adopted as a framework to classify cognitive processes in educational assessment, and organizing skills in a hierarchical structure where higher-order abilities depend on the mastery of lower-order ones [14, 15]. Despite its theoretical clarity, the practical application of BT remains challenging, as educators often struggle to consistently define and distinguish between cognitive levels due to their progressive dependencies and overlapping competencies [14]. The classification of questions is also inherently subjective since learners may interpret the same item differently depending on their knowledge, confidence, and problem-solving strategies [18, 29].

These challenges are particularly evident in the context of MCQs, one of the most widely used assessment formats in education [16, 23]. While MCQs are effective in evaluating factual knowledge (*Remembering*) and even higher cognitive skills, research consistently shows that they predominantly target the lowest levels of BT [23]. Although well-designed MCQs may assess intermediate levels such as *Applying* and *Analyzing*, the highest cognitive processes, of *Evaluation* and *Creation*, are rarely captured, as they typically require open-ended responses [15]. Moreover, even within the same question type of MCQs, items can vary substantially in complexity, a distinction that broad taxonomic levels can fail to capture [22]. To address these limitations, several works adopt a simplified representation of BT by grouping levels into lower-order and higher-order categories, thereby improving results and better reflecting the practical constraints of MCQ-based assessment [19, 31].

Classification using Bloom's Taxonomy

Early NLP approaches for BT classification relied on keyword/rule-based methods, using simple pre-processing techniques and predefined verb lists, or n-gram classifiers [8]. While achieving reasonable performance on lower levels, such as *Remembering* (around 75% accuracy), they perform poorly on higher levels, reflecting their inability to capture deeper semantic meaning [4, 20]. Building upon these limitations, more advanced approaches have explored emerging Machine Learning (ML) and NLP techniques to better capture semantic and contextual information.

In non-MCQ settings, transformer-based models such as BERT have demonstrated superior performance compared to traditional methods, significantly improving classification metrics across Bloom levels and effectively modeling contextual dependencies in educational texts [14, 21]. Additionally, studies in other languages and domains, such as Indonesian question classification, have combined lexical and syntactic feature extraction with ML models, including SVMs. These studies demonstrate that BT classification can generalize beyond English, although performance remains dependent on feature quality, POS-tagging-based feature extraction, and dataset size [13, 14]. These findings suggest that, while BT captures cognitive processes that are largely language independent, its classification remains inherently complex across linguistic settings. To address possible language-related limitations, other NLP studies, involving sentiment analysis or semantic similarity, have explored translation strategies, often achieving consistent or even improved performance over the native language [26].

Focusing on MCQ-based contexts, which introduce additional structural complexity, based on shorter question lengths, recent works have focused on leveraging transformer architectures with more sophisticated input representations to model relationships between question stems and answer options (e.g., using separated inputs such as "`[SEP] Options`" tokens) [24]. BERT models have been applied to classify MCQs according to BT, showing good performance, particularly at lower cognitive levels. Reported results indicate that classification accuracy improves significantly when simplifying the label space, increasing from 59.2% in the full six-level setting to 68.52% when merging underrepresented categories, and reaching 82.61% when further removing these classes [32]. These improvements are largely driven by the aggregation of less frequent classes, particularly the *Applying*, which is merged with higher-order categories to mitigate class imbalance and improve model learning. Similar trends are observed in Naive Bayes classifiers, where collapsing all of the BT levels, except *Remembering*, also leads to improved performance, albeit with limitations in cross-domain generalization [19]. Overall, these approaches are constrained by the imbalance and relatively small size of available datasets, which hinder the learning of higher-order cognitive distinctions [5]. Although performance improvements of the model understanding often come at the cost of reduced label granularity, as simplifying higher-order levels limits the ability to fully represent BT cognitive hierarchy [19, 32].

Even more recent approaches explore Large Language Models (LLMs) and specialized architectures for MCQ analysis, achieving strong performance on factual tasks, but still struggle with higher-order reasoning, with performance decreasing with cognitive complexity and item difficulty [2]. While these methods are not directly focused on BT classification, but rather on their generation, they provide relevant insights into how models handle different cognitive levels. In particular, these LLMs can generate and classify Bloom questions, but ensuring consistent quality and alignment across cognitive levels remains challenging, with discrepancies observed between automated and human evaluation [12]. Furthermore, as mentioned previously, the MCQ format itself introduces additional limitations to these strategies, as its structure can obscure item discrimination and hinder the reliable assessment of higher-order cognitive processes [9, 23].

3 Methodology

This section describes the study's pipeline, including the datasets, experimental setup, hyperparameter configurations, and evaluation metrics for the proposed models.

3.1 Data

In this study, we utilized two publicly available educational datasets, namely EduQG[2] and EduQuest[3] [7, 10]. From the latter, we primarily leveraged questions sourced from the OpenStax platform, covering Economics (3e), Microeconomics (2e and 3e), and Organizational Behavior. In contrast, EduQG does not provide explicit domain labels, but mainly contains science-related questions. Overall, merging these sources results in a more diverse dataset, both in terms of subject coverage and question structure.

To prepare these datasets, we first filtered only MCQs, which were identified either through an existing boolean indicator provided in the datasets or programmatically by parsing Question+Options structures in the text column. Then, only instances annotated

[2] `https://github.com/EchizenKurage/EduQG`
[3] `https://github.com/aegonwolf/EduQuest`

Table 1 Processed dataset summary statistics for each Bloom cognitive level (the last 3 columns represent a mean and standard deviation for each Bloom level).

BT Level	Count (%)	OPTs < 3 words	Fillblank ratio	Question char	OPTs char	OPTs words
1 – Remembering	1132 (26.88%)	516	49.29%	92 ± 51	103 ± 113	15 ± 17
2 – Understanding	1737 (41.25%)	372	45.83%	119 ± 70	156 ± 127	23 ± 20
3 – Applying	335 (7.96%)	147	27.76%	206 ± 167	122 ± 147	19 ± 24
4 – Analyzing	1007 (23.91%)	258	24.73%	214 ± 172	156 ± 119	25 ± 19

with a BT cognitive level were considered, ranging originally from 1 (*Remembering*) to 6 (*Creating*). In addition, to ensure greater consistency and mitigate class imbalance, only the first four levels of BT were considered. This decision was motivated mainly by: i) the limited representation of these classes in the merged dataset (166 and 16 instances out of a total of 4393 for *Evaluating* and *Creating* levels, respectively), and ii) the difficulty of identifying the top two levels of high-order thinking within the context of MCQs [15].

The merged dataset differs in both structure and domain coverage, encompassing variations in subject matter, as well as in question formats, including both standard and "fill-in-blank" MCQ items. This heterogeneity introduces increased complexity to the task while also supporting the research objective of evaluating the model's global performance in a cross-domain setting. Table 1 summarizes the main characteristics of the resulting dataset across the four BT levels considered in this study.

As shown in Table 1, the distribution of instances is skewed toward lower cognitive levels (*Remembering* and *Understanding*), which together comprise the majority of the dataset, while *Applying* is notably underrepresented. This discrepancy may stem from the annotator's subjective association of questions with BT levels, which can be inherently debatable. Despite the imbalance in sample sizes, some trends emerge, particularly structural differences in answer options. Lower levels, especially *Remembering*, show more options with fewer than three words, indicating a reliance on short, direct answers. However, this can be domain-dependent, as in questions involving numerical or formula-based reasoning, options often remain concise regardless of level, typically representing final results, which can bias length averages. Additionally, *Understanding* and *Analyzing* exhibit similar option lengths, suggesting that length alone does not capture cognitive complexity, while question length generally increases with higher levels.

The ratio of fill-in-blank items also decreases progressively with increasing cognitive level, indicating that higher-level questions rely less on simple recall formats.

3.2 Experimental setup

This section describes the experimental setup, including model configurations, input representations, and training procedures for the level-specific Bloom classification task.

As an initial approach, we implemented a solution based on predefined keyword/verb lists associated with each BT level (Table 2). Each question is assigned to a cognitive level based on the presence of specific keywords in the text. However, some action verbs, such as "classify", may appear at multiple levels, introducing ambiguity into this method, which requires certain heuristics. Prior to keyword matching, we applied simple text preprocessing to both the BT keyword lists and the question texts, including lowercasing, stopword removal, and stemming, in order to reduce lexical variability and improve matching consistency.

Table 2 BT levels and typical keywords of each Bloom level (adapted from [3]).

Bloom #	Bloom Level	Typical Keywords
1	Remembering	choose, define, find, how, label, list, match, name, omit, recall, relate, select, show, spell, tell, what, when, where, which, who, why
2	Understanding	classify, compare, contrast, demonstrate, explain, extend, illustrate, infer, interpret, outline, relate, rephrase, show, summarize, translate, understand
3	Applying	apply, build, choose, construct, develop, experiment, with, identify, interview, make, use, of, model, organize, plan, select, solve, utilize
4	Analyzing	analyze, assume, categorize, classify, compare, conclusion, contrast, discover, dissect, distinguish, divide, examine, function, inference, inspect, list, motive, relationships, simplify, survey, take, part, in, test, for, theme

However, this approach has several limitations, as it is highly sensitive to ambiguity, with the same keyword often corresponding to multiple cognitive levels; questions may contain a variety of relevant keywords or none at all, making the decision less straightforward. In such cases, tie-breaking rules are required, introducing heuristics that may not generalize well. Despite these limitations, this method provides a simple and interpretable baseline for comparison with the fine-tuned BERT-based models, although its performance is expected to be very limited.

Building upon the baseline approach, we explore transformer-based models to capture contextual and semantic information, with the main objective not only to evaluate global performance but also to assess the model's ability to distinguish between consecutive cognitive levels and to analyze disparities across languages. In particular, we aim to investigate whether models trained on the same data but directly translated into Brazilian Portuguese exhibit significant differences compared to their original form (English).

To this end, we employed two variants of pre-trained models for each language: i) a BERT-base version, and ii) a ModernBERT version [30]. For English, a standard BERT-base model and a ModernBERT model were used, each with 110M and 149M parameters, respectively. For Portuguese, we used the base version of BERTimbau (110M parameters) and a multilingual version of ModernBERT to assess the language [28, 17]. For comparison, the smaller multilingual ModernBERT model was chosen to ensure a similar number of parameters (140M). Regarding these Portuguese configurations, we obtained the translated version of the dataset using the `deep-translator` package[4].

To ensure comparability across these four models running process, the same data split is used for both versions, with an 80/20 train–test ratio and stratification over the target labels. A fixed random seed is applied so that predictions on the test set remain coherent across runs of these different models.

The input representation for all models was unified and standardized across all experiments to reduce noise and redundant information. Each instance is encoded as a sequence combining

[4] `https://pypi.org/project/deep-translator/`

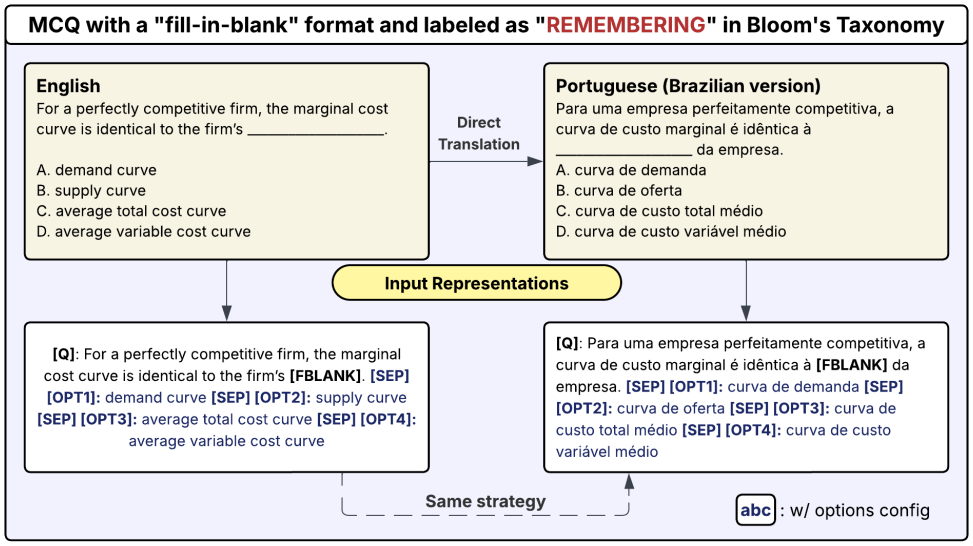


Figure 1 Example of input representations for the original English version and the Brazilian Portuguese translation.

the question and, optionally, its answer options, with the following structure: "`[Q] <question text>`" (with `[FBLANK]` tokens replacing blanks, when applicable), followed by "`[SEP]`" and the options starting as "`[OPT1] <option 1 text>`". These representation and masking techniques are applied consistently to both English and Portuguese inputs, as illustrated in Figure 1, through a fill-in-blank *Remembering* MCQ example. Based on this formulation, we explicitly control the inclusion of answer options in the input sequence. Therefore, all configurations are evaluated under two conditions: with and without answer options, enabling a direct assessment of their impact on the model's ability to infer the specific Bloom level.

BERT Hyperparameter tuning

Hyperparameter tuning was performed through a randomized search over predefined ranges for key parameters, including learning rate, weight decay, batch size, and warmup steps. To systematically explore this search space, we employed Optuna [1], running a total of 24 trials to identify the most suitable configurations. The optimization objective was the weighted F1-score, and the final hyperparameter settings were selected based on the 3 best-performing trials. So, the resulting configuration of hyperparameters used for all models in this study was essentially: `learning rate` $= 1.25 \times 10^{-5}$, `training and evaluation batch size` $=$ 16, `number of training epochs` $= 10$ (with a callback of early stopping using a patience of 2 epochs), `weight decay` $= 0.04$, and `warmup steps` $= 0.1$.

Class Selection

A complementary study was conducted to evaluate the impact of different class configurations on model performance. Initially, multiple setups were explored within a single model (BERT-base English), including merging adjacent levels (e.g., combining levels: $\{3, 4\}$ or $\{3, 4, 5, 6\}$), removing higher levels ($\{3, 4\}$ or $\{3, 5, 6\}$), or even both (e.g., dropping 5,6 and merging 3,4) [32, 19]. While some of these configurations yielded improved performance (primarily those

that reduced class granularity), the results are not directly comparable due to changes in the original classification task itself. Despite these improvements, the final decision was to maintain the four-level setting (dropping the 5-*Evaluation* and 6-*Creation* questions), as justified with the support of Table 1 and related work [19, 32]. Overall, evaluating each level independently was considered more meaningful for addressing RQ3 than merging classes, even at the cost of lower performance.

3.3 Evaluation metrics

To evaluate the performance of the fine-tuned models, we consider a set of standard classification metrics: accuracy, macro F1-score, weighted F1-score, and macro AUC. Accuracy provides a general measure of correctness but may be misleading in the presence of class imbalance. To address this, we report the macro F1-score, which treats all classes equally, highlighting performance on underrepresented classes, and the weighted F1-score that accounts for class frequency, in order to provide a more representative measure of overall performance. Given this class imbalance and to account for support, the weighted F1-Score is used as the primary metric for monitoring training and selecting the best model across epochs.

Beyond these traditional metrics, we also introduce additional evaluation measures, based on accuracy, to capture the ordinal nature of Bloom's Taxonomy levels: Adjacent Accuracy and H/L Order Accuracy. Since misclassifications between adjacent levels are less severe than those across distant levels, these metrics provide a more nuanced and interpretable assessment of the traditional accuracy by accounting for levels within adjacent classes or subgroups.

Adjacent Accuracy ($\text{Adj}_{\pm 1}$) considers a prediction correct if it matches the true label or differs by at most one adjacent class level, as shown in Equation 1.

$$\text{Adj}_{\pm 1} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (|\hat{y}_i - y_i| \leq 1) \tag{1}$$

In turn, H/L order accuracy evaluates whether predictions fall within the same ordinal group, being these groups: *Low-order* ($\{1, 2\}$) and *High-order* ($\{3, 4\}$). A prediction is considered correct if both predicted and true labels belong to the same group:

$$\text{H/L} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (g(\hat{y}_i) = g(y_i)), \quad g(y) = \begin{cases} 0, & \text{if } y \in \{1, 2\} \text{ (Low)} \\ 1, & \text{if } y \in \{3, 4\} \text{ (High)} \end{cases} \tag{2}$$

This metric can be generalized to the full six-level Bloom taxonomy by redefining the grouping function accordingly. However, in our four-level setting, it behaves similarly to adjacent accuracy ($\text{Adj}_{\pm 1}$), effectively ignoring only confusions between the intermediate *Understanding* and *Applying* levels to focus on broader cognitive distinctions instead. In essence, this corresponds to a dichotomization ($\{1,2\}$ and $\{3,4\}$) of the resulting confusion matrix after training the model on the four-level classification task, rather than explicitly training a separate model for high/low-order classification; this would constitute a simpler task, as the model would not need to differentiate between all classes and would likely yield higher accuracy values in the binary classification than this new metric in the multi-class task.

Finally, in the evaluation phase, we explore two alternative strategies for interpreting the model output probabilities. The first corresponds to the standard *argmax* approach, where the predicted class is selected as the one with the highest probability, while the

second approach computes a *weighted average* of the class probabilities, leveraging the ordinal structure of the labels. This method, which led to more robust models in other NLP tasks [25], assigns greater weight to lower-probability classes, particularly at levels $\{3, 4\}$, which enables smoother predictions and provides additional insight into borderline cases where the model expresses uncertainty.

Besides, to reduce the impact of randomness during training, each experiment is repeated 5 times with different random seeds. The reported results correspond to the mean of the evaluation metrics across these runs, providing a more reliable estimate of model performance. Considering all configurations, this yields a total of 40 training runs, derived from 4 model variants (with the corresponding language), 2 input settings (with and without answer options), and 5 repetitions per configuration. The two methods shown to evaluate the probability vector don't affect the entire training (across the 40 training runs mentioned), but only the metrics and confusion matrix derived from the predictions.

4 Results

This section begins with an evaluation of the baseline model, keyword-based approach, followed by a comprehensive analysis of the BERT-based configurations. Multiple experimental setups are compared, with particular focus on their respective advantages and the identification of consistent performance patterns across models, input representations, and prediction strategies. Subsequently, a targeted error analysis is conducted to examine common misclassification trends, with special attention given to discrepancies observed between English and Portuguese predictions, supported by illustrative examples. Finally, the analysis is extended through the use of profiling variables, enabling a more granular assessment of model performance across Bloom levels, as well as across structural and contextual dimensions, such as fill-in-blank questions and topic-related categorical features.

Baseline Model

Before advancing to the evaluation of the learning-based models, it is important to analyze the performance of the baseline approach. First, it is relevant to notice that the baseline classifier operates deterministically over the entire dataset, as it does not involve any training phase, but instead assigns labels based on predefined keyword rules, already shown in Table 2. The resulting confusion matrix is shown in Figure 2.

The baseline classifier yields notably poor performance, as reflected in the low metric values (all close to 0.2). These results highlight the limitations of relying solely on keyword-based heuristics for Bloom level classification. From the confusion matrix, it is evident that the classifier exhibits a strong bias toward specific classes, with emphasis on *Applying* and, to a lesser extent, *Remembering*, which are frequently "overpredicted" regardless of the true label. For instance, a large portion of *Understanding* instances are incorrectly assigned to *Applying*, indicating that the presence of certain keywords may be misleading or overly generalized across levels.

Moreover, a substantial portion of the dataset remains unclassified, as many questions do not contain any of the predefined keywords. This is reflected in the large number of samples in the "no_keyword" category (+80% of the samples in {1,2,3} levels and $\approx$ 74% in the 4th level - *Analyzing*), which significantly limits the coverage and effectiveness of this approach.

These results reinforce the limitations of these simple rule-based methods, underscoring the need for models that capture deeper contextual and semantic information.

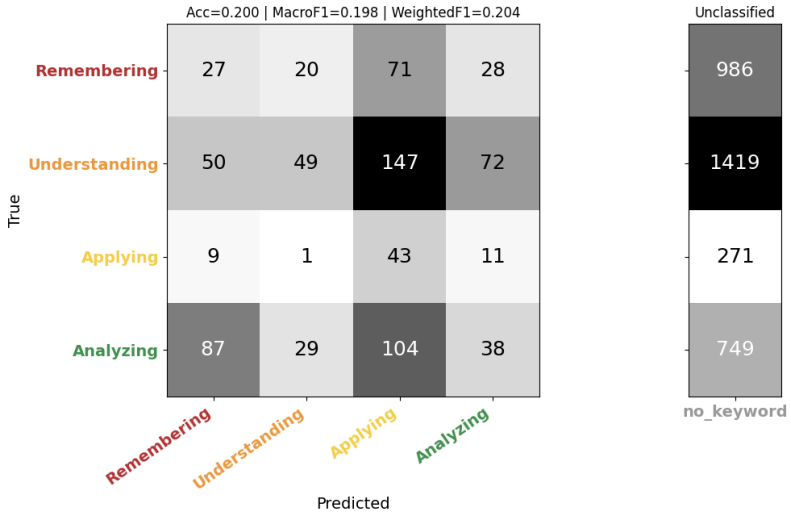


Figure 2 Confusion Matrix of the Baseline Model: Classification by keyword-based approach.

Transformer-based models

Moving on to the fine-tuned BERT models, Table 3 presents the performance of the different BERT-based configurations across languages, input settings, and prediction strategies. In this table, the colored arrows indicate the comparison by excluding the options in one color and changing the prediction strategy in another color. In turn, the arrows in ModernBERT compare each line to its mirror in the "simpler" BERT, without this clear color separation between options and strategy subgroups.

Table 3 BERT-Models configurations results for the specific class of BT: mean after 5 runs of each config.

Model Configuration w/ the 1st 4 classes of BT				Classification Metrics (mean of 5 runs)					
Model	Lang.	+ OPT?	Pred. Strat.	Acc.	MacroF1	WeightedF1	Adj (± 1)	H/L order	Macro AUC
BERT-base	EN	✓	argmax	**0.6989**	0.6243	**0.6922**	0.8598	**0.8273**	**0.8665**
BERTimbau-base	PT	✓	argmax	0.6892	0.6201	0.6824	0.8486	0.8202	0.8635
BERT-base	EN	✗	argmax	0.6968 ↓	0.6209 ↓	0.6879 ↓	0.8520 ↓	0.8256 ↓	0.8625 ↓
BERTimbau-base	PT	✗	argmax	0.6878 ↓	**0.6361** ↑	0.6837 ↓	0.8475 ↓	0.8187 ↓	0.8566 ↓
BERT-base	EN	✓	weighted avg	0.6486 ↓	0.5882 ↓	0.6613 ↓	**0.8982** ↑	0.8244 ↓	**0.8665**
BERTimbau-base	PT	✓	weighted avg	0.6372 ↓	0.5754 ↓	0.6482 ↓	0.8887 ↑	0.8187 ↓	0.8635
BERT-base	EN	✗	weighted avg	0.6377 ↓↓	0.5780 ↓↓	0.6491 ↓↓	0.8963 ↓↑	0.8223 ↓↓	0.8625 ↓
BERTimbau-base	PT	✗	weighted avg	0.6489 ↑↓	0.5893 ↑↓	0.6552 ↑↓	0.8816 ↓↑	0.8211 ↑↑	0.8566 ↓
ModernBERT	EN	✓	argmax	**0.6904** ↓	**0.6292** ↑	**0.6884** ↓	0.8550 ↓	0.8199 ↓	**0.8686** ↑
mModern-small	PT	✓	argmax	0.6878 ↓	0.6224 ↑	0.6853 ↓	0.8503 ↑	0.8199 ↓	0.8590 ↓
ModernBERT	EN	✗	argmax	0.6871 ↓	0.6270 ↑	0.6838 ↓	0.8529 ↑	**0.8230** ↓	0.8672 ↓
mModern-small	PT	✗	argmax	0.6840 ↓	0.6191 ↓	0.6793 ↓	0.8531 ↑	0.8202 ↑	0.8563 ↓
ModernBERT	EN	✓	weighted avg	0.6230 ↓	0.5718 ↓	0.6412 ↓	**0.9006** ↑	0.8166 ↓	**0.8686** ↑
mModern-small	PT	✓	weighted avg	0.6453 ↑	0.5889 ↑	0.6579 ↑	0.8859 ↓	0.8178 ↓	0.8590 ↓
ModernBERT	EN	✗	weighted avg	0.6380 ↑	0.5837 ↑	0.6502 ↑	0.8916 ↓	0.8214 ↓	0.8672 ↑
mModern-small	PT	✗	weighted avg	0.6313 ↓	0.5765 ↓	0.6441 ↓	0.8951 ↑	0.8192 ↓	0.8563 ↓

Overall, these transformer-based models outperform the baseline approach by a large margin, confirming the importance of contextual and semantic modeling for this task. This is also aligned with the nature of the annotation process itself, where human annotators rely

not only on lexical cues but also on semantic understanding, external knowledge, pattern recognition, and elements of critical thinking. While transformer models are capable of capturing contextual and semantic information, fully replicating the critical reasoning applied by human annotators remains a challenging aspect.

Across both BERT-base and ModernBERT variants, the results show very similar performance between the English and Portuguese models, in accuracy and weighted F1-score, above all. However, a slight and consistent performance drop is observed for the Portuguese versions. This can be attributed to the use of direct translation, which may introduce noise or subtle semantic inconsistencies, especially given the greater lexical variability of the Portuguese language. When comparing model architectures, ModernBERT exhibits behaviour similar to that of standard BERT-base models, although it does not consistently improve performance and, in some cases, yields slightly lower metric values. This suggests that the more recent architecture does not necessarily translate into better results for this specific task, possibly due to the dataset characteristics or task complexity/meticulousness.

Regarding input representation, the inclusion of answer options leads to marginal improvements in performance, particularly in accuracy and weighted F1-score. This indicates that while the additional context provided by the options is beneficial, its impact is relatively limited, as the question itself already contains most of the relevant information.

In terms of prediction strategy, the *weighted average* approach clearly outperforms the standard *argmax* method in Adjacent Accuracy (± 1), as expected given its design to account for ordinal proximity between classes. However, this comes at the cost of lower performance in strict classification metrics such as accuracy and F1-score, as well as slightly worse results for the H/L order metric. This highlights the trade-off between prioritizing exact level classification and ordinal consistency and reinforces our choice made on Section 3.2.

Macro AUC showed strong values and remained highly consistent across all configurations, with stable class separability regardless of the model, language, or input setup.

Thus, these results confirm that transformer-based approaches performed way better than the baseline classifier, while still revealing consistent patterns across configurations. We now proceed to a more detailed error analysis, focusing on the BERT-base English and Portuguese variations to better understand the differences in their prediction behavior.

Error analysis between languages

Figure 3 illustrates a side-by-side comparison of predictions for the English and Portuguese BERT-base models (with options tokenized) on the same set of test examples. While overall performance is similar, several instances reveal divergences between the two versions, where one model correctly predicts the Bloom level and the other fails, highlighting the impact of translation on semantic interpretation.

From the 30 samples, instances 7, 18, 20, 22, and 27 exhibit divergences between the English and Portuguese predictions, while instance 8 shows divergence in the prediction strategy for English, and instance 18 for Portuguese. To better illustrate these discrepancies, we revisit the example presented in Figure 1, which corresponds to example 27 in Figure 3. This is a MCQ where the model correctly predicts the original English version as *Remembering*, but classifies the Portuguese version as *Understanding*. This difference is also reflected in the predicted probabilities obtained: the English model assigns 0.56 to *Remembering* and 0.44 to *Understanding*, whereas the Portuguese model assigns 0.17 to *Remembering* and 0.82 to *Understanding*, with the remaining probability mass distributed across the other classes.

This example highlights a subtle cross-linguistic difference that can impact the model prediction. Notably, this example contains none of the main keywords typically used to associate a BT level (Table 2), making the classification more dependent on contextual

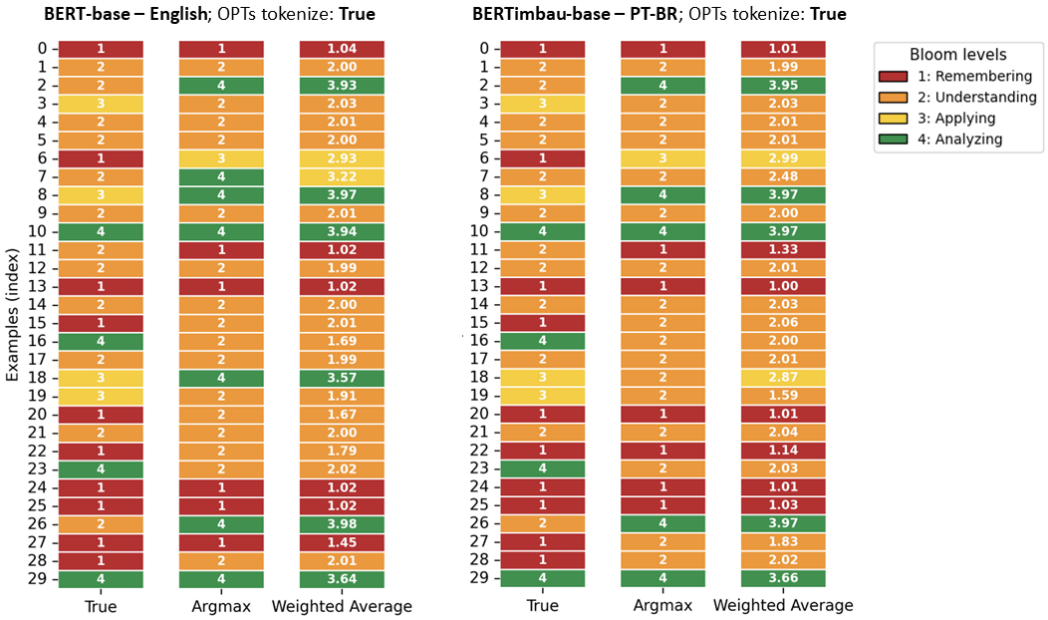


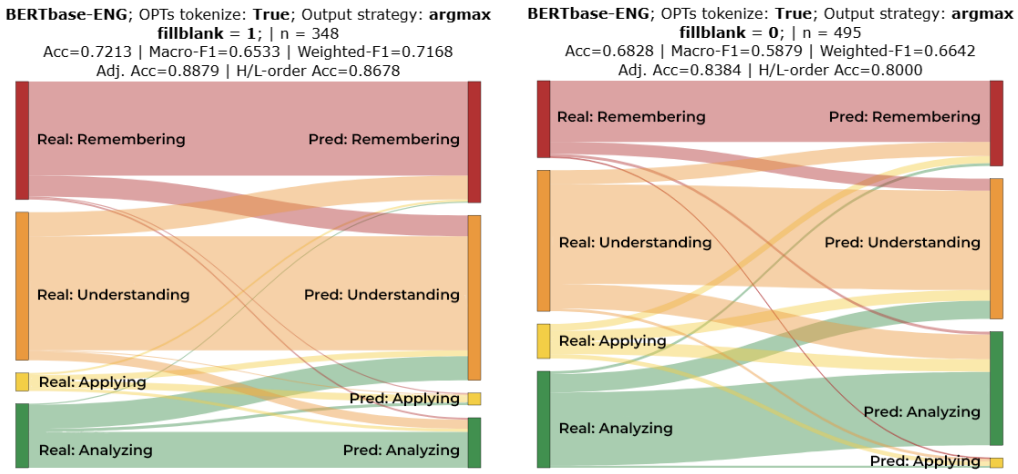
Figure 3 Model output strategies for each language best configuration: 30 examples of test set.

cues than on explicit lexical signals. The Portuguese formulation introduces a syntactic restructuring ("à [FBLANK] da empresa") that weakens the direct possessive cue present in English ("the firm's [FBLANK]"), potentially reducing the salience of the underlying economic relationship and making the cognitive level less distinguishable. So, this reformulation may shift the question toward a different classification, as the blank is no longer positioned at the end of the sentence, which may alter the meaning that the model derives from the question. Still, even in the English version, the prediction remains borderline, as reflected in the close probabilities between the two classes, likely due to slight nuances.

Variables Profiling

This analysis is conducted post hoc, by profiling the model predictions according to specific variables, in order to assess whether the model behaves differently across subsets of the data. This was performed for both the fill-in-blank format and the question subject source, indicated by a "Book" field in the original dataset.

For a single run of BERT-base English model with answer options and softmax strategy, performance across fill-in-blank and non-fill-in-blank questions remains largely comparable, with consistent variations in the evaluated metrics: Acc = (0.7213 vs 0.6828) and Weighted F1 = (0.7168 vs 0.6642). From the Sankey plots in Figure 4, most misclassifications occur between consecutive Bloom levels, which is consistent with the relatively high values observed for the Adjacent and H/L Order accuracy metrics. This suggests that predictions tend to remain close to the true class. However, *Applying* remains the most challenging class, often being misclassified with adjacent levels, with notable confusion between *Understanding* and *Analyzing*, likely because instances requiring *Analyzing* also involve *Understanding* (but not vice-versa), as well as similarities in question and answer length and structure. Also, it is important to note that the test set for this profiling variable is unbalanced, with a larger



(a) For "fill-in-blank" questions.

(b) For non "fill-in-blank" questions.

Figure 4 Sankey Plot and metrics comparison between type of questions (inspired by [12]).

proportion of non-fill-in-blank instances (495 vs 348), which may influence the evaluation, particularly for *Applying*, where the disparity in the number of instances is more pronounced compared to other classes.

Comparing both settings, the fill-in-blank case (Sub-figure 4a) shows slightly more concentrated flows along correct mappings, whereas the non-fill-in-blank case (Sub-figure 4b) exhibits marginally more dispersed flows, particularly between *Understanding*, *Applying*, and *Analyzing*, with notably 61 instances of misclassifications between *Understanding* and *Analyzing* classes. Despite the performance remaining similar across both settings, with only minor differences in metrics and flow distributions, the patterns evidenced may also be influenced by the disparity in the typology of questions.

Across topic-based groupings, the classification performance remains largely consistent, with variations of around 0.02 in the main evaluated metrics. An exception is observed in the "Organizational Behavior" category, where performance deviates more noticeably (a 0.15 variation), likely due to the limited number of samples available for this domain.

5 Conclusion

This study investigated the effectiveness of different approaches for classifying MCQs according to BT, focusing on contextual modeling, input structure, and cross-linguistic robustness. In line with the defined Research Questions (RQs), several key findings emerge.

Regarding RQ1, the results demonstrate that simple rule-based methods are insufficient to capture the complexity of this task, reinforcing the need for models that can benefit from semantic and contextual information. Transformer-based models, specifically BERT and ModernBERT variants, showed significant improvements over the baseline classifier, delivering consistent and robust performance across configurations, with satisfactory metric values. The inclusion of answer options provided marginal gains, suggesting that while additional context is beneficial, the question itself carries most of the discriminative signal. Furthermore, the comparison of prediction strategies highlighted a trade-off between exact classification and ordinal consistency, with the *weighted average* approach improving the created "proximity" metrics at the expense of standard accuracy.

For RQ2, the interlingual analysis revealed that models trained on translated Portuguese data achieve performance comparable to their English counterparts, although with a slight and consistent degradation. A small error analysis demonstrated that even subtle syntactic and structural differences introduced by translation can impact model predictions, particularly in borderline cases where class distinctions are less clear. These findings emphasize the sensitivity of such models to linguistic nuances and still highlight the challenges of maintaining semantic equivalence across languages.

Concerning RQ3, most errors occur between adjacent Bloom levels, particularly *Understanding*, *Applying*, and *Analyzing*, highlighting the difficulty of fine-grained distinctions. Overall, models perform better on *Remembering*, with performance decreasing at the three higher levels mentioned, which is consistent with the literature. This is further influenced by dataset limitations, especially the under-representation of the *Applying* class. Additionally, performance remains stable across formats and the majority of domains, with gains in weighted F1-score for fill-in-blank questions and minor deviations between consecutive levels. Overall, errors are mainly driven by class proximity, data imbalance, and the inherent ambiguity of Bloom-level distinctions.

Beyond these findings, some limitations should be considered. The dataset is heterogeneous and partially uncurated, with class imbalance and domain variability, particularly affecting the most underrepresented class. Moreover, the task itself is inherently complex, as BT categories are relatively coarse and often overlapping, making fine-grained distinctions ambiguous. Despite this, the models produce coherent and consistent predictions, with cross-linguistic results showing aligned behavior between English and translated Portuguese data. This indicates that transformer-based approaches capture underlying cognitive patterns that generalize across languages, although performance is still influenced by data quality and the limits of the taxonomy.

Future work may focus on improving both data quality and modeling approaches. On the data side, expanding the dataset with additional MCQs while ensuring a more balanced distribution across Bloom levels, topics, and question types would likely lead to more robust conclusions, as well as support further class-separation analyses, including simplified two-class settings. Additionally, incorporating complementary educational frameworks, such as Item Response Theory (IRT), could provide more nuanced estimates of question difficulty and enable joint modeling of cognitive level and complexity, though this would require student response data [27, 24]. Extending this idea from a modeling perspective, future work could explore hybrid approaches that integrate textual representations with external educational signals, such as IRT-based features or other metadata, to better capture both semantic and psychometric aspects of the task. The integration of external knowledge sources, potentially through retrieval-based mechanisms, may further enhance contextual understanding, while iterative refinement of such hybrid models could improve generalization and robustness. Building on such a pipeline, a promising direction would be the creation of a high-quality Portuguese dataset, for instance, by leveraging LLMs to generate and label MCQs according to BT or richer multi-label schemes, informed by both semantic and difficulty-related signals.

References

1 Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

2 Fernando R Altermatt, Andres Neyem, Nicolás I Sumonte, Ignacio Villagrán, Marcelo Mendoza, and Hector J Lacassie. Evaluating the performance of large language models on the conacem anesthesiology certification exam: A comparison with human participants. *Applied Sciences*, 15(11):6245, 2025.

3 Lorin W Anderson and David R Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.

4 Wen-Chih Chang and Ming-Shun Chung. Automatic applying bloom's taxonomy to classify and analysis the cognition level of english question items. In *2009 Joint Conferences on Pervasive Computing (JCPC)*, pages 727–734, 2009. doi:10.1109/JCPC.2009.5420087.

5 Syaamantak Das, Shyamal Kumar Das Mandal, and Anupam Basu. Identification of cognitive learning complexity of assessment questions using multi-class text classification. *Contemporary Educational Technology*, 12(2):ep275, 2020.

6 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar 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. doi:10.18653/v1/N19-1423.

7 Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. Eduqg: A multi-format multiple-choice dataset for the educational domain. *IEEE Access*, 11:20885–20896, 2023. doi:10.1109/ACCESS.2023.3248790.

8 S.S. Haris and Nazlia Omar. Bloom's taxonomy question categorization using rules and n-gram approach. *Journal of Theoretical and Applied Information Technology*, 76:401–407, June 2015.

9 Kirk Hillsley. Question format is the best predictor of item discrimination: a multivariable analysis. *Journal of Microbiology and Biology Education*, 26(3):e00205–25, 2025.

10 Oliver Holl, Filipe Szolnoky Cunha, David Streuli, and TimothÃ© Laborie. Eduquest: Lecture texts and questions for higher education. In Benjamin PaaÃŸen and Carrie Demmans Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 849–856, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society. doi:10.5281/zenodo.12729970.

11 Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984, 2018. doi:10.1016/j.ipm.2018.06.007.

12 Kevin Hwang, Kenneth Wang, Maryam Alomair, Fow-Sen Choa, and Lujie Karen Chen. Towards automated multiple choice question generation and evaluation: aligning with bloom's taxonomy. In *International Conference on Artificial Intelligence in Education*, pages 389–396. Springer, 2024. doi:10.1007/978-3-031-64299-9_35.

13 Selvia Ferdiana Kusuma, Daniel Siahaan, and Umi Laili Yuhana. Automatic indonesia's questions classification based on bloom's taxonomy using natural language processing a preliminary study. In *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 1–6. IEEE, 2015.

14 Yuheng Li, Mladen Rakovic, Boon Xin Poh, Dragan Gasevic, and Guanliang Chen. Automatic classification of learning objectives based on bloom's taxonomy. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 530–537. International Educational Data Mining Society, July 2022. doi:10.5281/zenodo.6853191.

15 Qian Liu, Navé Wald, Chandima Daskon, and Tony Harland. Multiple-choice questions (mcqs) for higher-order cognition: Perspectives of university teachers. *Innovations in Education and Teaching International*, 61(4):802–814, 2024. doi:10.1080/14703297.2023.2222715.

16 Kseniia Marcq, Erika Jassuly Chalén Donayre, and Johan Braeken. The role of item format in the pisa 2018 mathematics literacy assessment: A cross-country study. *Studies in Educational Evaluation*, 83:101401, 2024. doi:10.1016/j.stueduc.2024.101401.

17 Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. mmbert: A modern multilingual encoder with annealed language learning, 2025. doi:10.48550/arXiv.2509.06888.

18 Susana Masapanta-Carrión and J. Ángel Velázquez-Iturbide. A systematic review of the use of bloom's taxonomy in computer science education. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education, SIGCSE 2018, Baltimore, MD, USA, February 21-24, 2018*, pages 441–446, February 2018. doi:10.1145/3159450.3159491.

19 Alan D. Mead and Chenxuan Zhou. Using machine learning to predict bloom's taxonomy level for certification exam items. *Journal of Applied Testing Technology*, 23:53–71, October 2022. URL: https://jattjournal.net/index.php/atp/article/view/170341.

20 Nazlia Omar, Syahidah Sufi Haris, Rosilah Hassan, Haslina Arshad, Masura Rahmat, Noor Faridatul Ainun Zainal, and Rozli Zulkifli. Automated analysis of exam questions according to bloom's taxonomy. *Procedia - Social and Behavioral Sciences*, 59:297–303, 2012. Universiti Kebangsaan Malaysia Teaching and Learning Congress 2011, Volume I, December 17 – 20 2011, Pulau Pinang MALAYSIA. doi:10.1016/j.sbspro.2012.09.278.

21 Dhaval Patel, Krish Bhikadiya, Nikita Bhatt, Ronakkumar Patel, and Trusha Patel. Impact of bert on evaluating the quality of question papers using bloom's taxonomy. In *International Conference on ICT for Sustainable Development*, pages 65–74. Springer, 2024.

22 Radek Pelánek, Tomáš Effenberger, and Jaroslav Čechák. Complexity and difficulty of items in learning systems. *International Journal of Artificial Intelligence in Education*, 32(1):196–232, 2022. doi:10.1007/s40593-021-00252-4.

23 Hamdollah Ravand, Reza Shahi, Farshad Effatpanah, and Ali Moghadamzadeh. Measuring cognitive levels in high-stakes testing: A cdm analysis of a university entrance examination using bloom's taxonomy. *Frontiers in Education*, Volume 10 - 2025, 2025. doi:10.3389/feduc.2025.1644811.

24 M. Ravikiran, T. Sharma, A. Bhavsar, and R. Saluja. Gisa: Gradual information selection attention for mcq difficulty estimation. *Lecture Notes in Computer Science*, 15877:193–206, 2025. doi:10.1007/978-3-031-98414-3_14.

25 Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 551–557, 2024. URL: https://aclanthology.org/2024.propor-1.59/.

26 Ruan Chaves Rodrigues, Jéssica Rodrigues da Silva, Pedro Vitor Quinta de Castro, Nádia Félix Felipe da Silva, and Anderson da Silva Soares. Multilingual transformer ensembles for portuguese natural language tasks. In *Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese co-located with XII Symposium in Information and Human Language Technology (STIL 2019)*, pages 27–38, March 2020.

27 Duden Saepuzaman, Edi Istiyono, Heri Retnawati, et al. Characteristics of two-tier multiple choice (ttmc) high order thinking skill (hots) instruments and the estimation of hots abilities of vocational education students using the item response theory approach. *Journal of Engineering Education Transformations*, pages 13–22, 2025.

28 Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: Pretrained bert models for brazilian portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems*, pages 403–417, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-61377-8_28.

29 J. K. Stringer, Sally A. Santen, Eun Lee, Meagan Rawls, Jean Bailey, Alicia Richards, Robert A. Perera, and Diane Biskobing. Examining bloom's taxonomy in multiple choice questions: Students' approach to questions. *Medical Science Educator*, 31(4):1311–1317, 2021. doi:10.1007/s40670-021-01305-y.

30 Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. doi:10.18653/v1/2025.acl-long.127.

31 Lamine Zaidi, Karri Grob, Seetha Monrad, Joshua Kurtz, Andrew Tai, Asra Ahmed, Larry Gruppen, and Sally Santen. Pushing critical thinking skills with multiple-choice questions: Does bloom's taxonomy work? *Academic Medicine*, 93:1, December 2017. doi:10.1097/ACM.0000000000002087.

32 James Zhang, Casey Wong, Nasser Giacaman, and Andrew Luxton-Reilly. Automated classification of computing education questions using bloom's taxonomy. In *Proceedings of the 23rd Australasian computing education conference*, pages 58–65, 2021. doi:10.1145/3441636.3442305.