Department of Information and Computing Science

Applied Data Science Master Thesis

Automatic Annotation of Dutch Educational Assessment Questions using Large Language Models

First examiner:

Matthieu J. S. Brinkhuis

Second examiner:

Gerard Barkema

Candidate:

Isabela Harumi Lopes Motoki

In cooperation with:

Lientje Maas (Cito)

Zoë ten Napel (Cito)

June 30, 2025

Abstract

This study is aimed at the automatic evaluation of curriculum alignment. Curriculum alignment refers to the extent to which learning objectives, instructional activities, and assessments are coherently aligned. Traditionally, measuring this alignment is a time-consuming and often subjective process, since it typically involves evaluating all educational materials with the learning objectives of the curriculum. To address this, the research explores the use of large language models (LLMs) to automate the annotation of Dutch assessment questions with subject-specific concepts. Specifically, it investigates both generative (GPT-4.1 nano) and non-generative (mBERT) models using a labeled dataset of Dutch statistics questions. Results indicate that LLMs show strong potential in this domain: GPT achieved up to 71.1% accuracy and 62.2% macro F1 score, while mBERT reached 91.7% accuracy and 83.7% macro F1 score. Additionally, prompt engineering significantly enhances GPT's performance, leading to substantial gains. The findings also highlight the importance of careful adaptation and evaluation across diverse educational contexts and task types, as performance varied depending on question categories and subject matter. This research contributes to the integration of AI in education by providing an effective solution for question annotation and offering insights into which approaches are better suited for different educational scenarios. As a result, educators can better align assessments with learning objectives and enhance the overall learning experience.

Contents

1. Introduction

It is widely recognized that education has a lasting impact on people's lives. It shapes their opportunities, critical thinking, and ability to participate in society. Therefore, educational programs must be carefully planned to promote effective learning experiences. A crucial part of this planning process is having a well-structured curriculum that clearly defines the learning objectives, the instructional activities and supporting materials, and the assessments (Anderson, 2002). As a result, when these three components are aligned, they enable both educators and students to engage in purposeful educational activities, fostering meaningful learning. This research, therefore, aims to develop an automatic evaluation of curriculum alignment, contributing to the design and improvement of educational programs and ultimately promoting effective learning.

Despite its importance, ensuring curriculum alignment is not a trivial task. Verifying, for example, to what extent an assessment measures the intended curricular objectives often requires linguistic proficiency as well as a deep understanding of pedagogical goals (Nappi, 2017). Traditionally, this process has relied on manual analyses by educational experts. However, such methods are often time-consuming, resource-intensive, and typically involve collaboration among multiple specialists. Moreover, manual methods are prone to subjectivity, potentially leading to omissions or overly generalized interpretations, thereby reducing the effectiveness of this approach.

These limitations have become increasingly evident with the exponential growth of online educational content, rendering conventional manual approaches insufficient to handle large volumes of educational data. As a result, there is growing interest in developing automated solutions that support educators in organizing and evaluating educational content more efficiently while preserving alignment with learning objectives.

Considering that educational institutions typically evaluate students' level of comprehension through exams, the questions used in these assessments play a crucial role in the learning process. Consequently, a common approach to evaluate curriculum alignment involves analyzing whether assessment questions adequately reflect the intended learning objectives. This process typically involves annotating each item according to pedagogical labels, such as subject-specific concepts, cognitive processes, or skill levels, and therefore enables educators assess how well educational content aligns with formal curricula.

To address this task, various automated solutions employing machine learning (ML) and Natural Language Processing (NLP) techniques have been proposed (T. Huang et al., 2023; Li et al., 2024; Osman & Yahya, 2016; Tian et al., 2022). Early methods commonly used text embedding techniques and traditional machine learning classifiers, but they often neglected the crucial relationship between question and underlying subject concepts, resulting in suboptimal annotation performance.

As a consequence, more recent studies have been focusing on NLP techniques, especially those involving large language models (LLMs), characterized by extensive pretraining on vast textual corpora. These models have demonstrated significant improvement in classification accuracy by leveraging pretrained language models and incorporating external information (Li et al., 2024). Furthermore, the use of generative models, such as Generative Pre-trained Transformer (GPT), has represented a paradigm shift, gaining widespread attention for their performance across various NLP tasks (Faraby et al., 2024).

Therefore, in order to automate question annotation and hence support curriculum alignment applications, this research lies in exploring the capability of LLMs to understand and classify educational questions. Moreover, given that one of the key advantages of using LLMs is their ability to perform zero-shot and few-shot learning, this study aims to understand how analysts can reduce the dependence on large and task-specific labeled datasets for model fine-tuning. This approach enhances the scalability and adaptability of automated annotation methods, making them suitable for diverse educational purposes and environments.

It is also important to note that the majority of existing research has focused predominantly on English-language educational content, with comparatively fewer studies addressing other languages, such as Dutch. This linguistic limitation motivates the present study, which aims to conduct a thorough evaluation of LLMs' effectiveness in annotating Dutch educational questions.

Specifically, the research questions addressed in this paper are as follows:

- **RQ1:** *How well can large language models annotate Dutch assessment questions with curriculum subject-specific concepts to support their alignment?*

  By understanding the performance of LLMs in annotating questions with subject-specific concepts, it is possible to assess their applicability and reliability in evaluating educational content. This knowledge is important for developing systems that can automatically annotate questions and potentially enhance diverse curriculum alignment approaches.

Additionally, the following subquestions guide the research:

- **SQ1:** *What is the impact of prompt format and the provided curricular context information on the annotation performance of generative models?*

  Considering that prompt engineering plays a key role in shaping the output of generative models, analyzing different instructions or adding context information, such as the course and academic level associated with each question, helps identify strategies to improve annotation accuracy.

- **SQ2:** *In what ways do generative and non-generative multilingual models differ in their accuracy and robustness at annotating Dutch assessment questions by curriculum subject-specific concepts?*

  By examining how these different model types perform, the study seeks to highlight their respective strengths and limitations. This comparison provides insights into which approaches are more reliable, in particular in multilingual or low-resource educational settings.

In summary, the overall goal of this research is the enhancement of educational content evaluation through automation. By integrating automatic question annotation, educators can ensure the use of questions aligned with learning objectives, enhancing the learning experience. This capability allows educators to focus more on deeper cognitive engagement instead of spending time crafting questions.

2. Literature review

This literature review provides an overview of the key developments in automated educational question annotation. It traces the evolution from traditional machine learning approaches to modern transformer-based models, highlighting improvements in performance and efficiency. Additionally, it discusses the growing role of prompt engineering in maximizing the effectiveness of LLMs and explores multilingual applications, with a focus on challenges and strategies for low-resource languages such as Dutch. Together, these sections establish the foundation for understanding the technical and contextual landscape that motivates this study.

2.1 Evolution of Question Classification Methods

Research on automated question annotation initially focused on traditional ML techniques. Early studies predominantly utilized algorithms such as Support Vector Machines (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression, often combined with feature extraction methods like Bag-of-Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF), and Part-of-Speech (POS) tagging. For example, Silva et al. (2018) reviewed 80 studies on automatic question classifiers and found SVM to be the most common ML algorithm, with BOW and TF-IDF as primary feature extraction techniques. When applied properly and carefully, such methods have demonstrated strong performance; Osman and Yahya (2016), for instance, achieved 76% accuracy with SVM when classifying assessment questions according to cognitive levels.

Subsequent research started exploring methods such as word embeddings and neural networks, enabling models to improve classification performance without relying solely on traditional feature engineering. Tian et al. (2022), for example, used word embedding-based methods and achieved a 79% F1-score and 79.5% accuracy in question classification, outperforming classic ML methods like XGBoost and LSTM. Similarly, Mustafidah et al. (2022) used POS tagging combined with NLP techniques to classify questions into cognitive levels with 82% accuracy.

Despite these advancements, studies that focused on these traditional ML or deep learning methods demonstrated some challenges and limitations. Besides their dependence on using proper engineered features, they also require large labeled datasets to achieve reasonable accuracy, which are not always available in educational contexts.

The emergence of LLMs has shifted much of the recent research focus toward pretrained architectures. Unlike traditional ML models, LLMs benefit from extensive pretraining on massive corpora, enabling them to capture complex contextual and semantic patterns with less reliance on task-specific annotated data. This shift has substantially reduced the data requirements previously needed for effective model fine-tuning. For instance, pretrained models like BERT have been applied successfully in tasks such as taxonomy-based question annotation. Zemlyanskiy et al. (2021) pretrained BERT to jointly predict words and entities as movie tags from reviews, while T. Huang et al. (2023) proposed improvements to BERT for annotating subject-specific concepts in educational questions and solutions. Moreover, Li et al., 2024 leveraged generative models such as GPT, Llama2, and Mixtral, achieving relevant results in classifying math questions.

In summary, the evolution from classical ML models to LLMs reflects a trend towards leveraging richer contextual representations and larger data sources. While ML relies on large training data and encounters challenges when generalizing across diverse educational contexts, pretrained models, especially generative LLMs, demonstrate promising capabilities in zero-shot and few-shot learning scenarios, enabling effective question classification even in low-resource or multilingual environments.

2.2 LLMs in Educational Context

Large generative models such as GPT, Llama, or Gemini have been demonstrating strong zero-shot or few-shot learning capabilities, significantly reducing the dependence on task-specific annotated datasets. For instance, Moore et al. (2024) used GPT-4 to generate approximately 40 concept classes for multiple-choice questions using only the question text. The results, compared to human-labeled gold standards, achieved 56% accuracy, which shows the model's ability to generalize given unsupervised conditions. Similarly, Li et al. (2024) introduced a method applying LLM zero-shot and few-shot learning to link math questions with relevant subject-specific concepts without requiring external solution texts or extensive fine-tuning, achieving 90% accuracy with GPT-4.

These models' power stems from their training on vast and diverse real-world datasets, granting them universal language understanding and generation capabilities. As the use of LLMs has become widespread in educational question classification, research has increasingly focused on optimizing prompt design to leverage their learning capabilities. Studies highlight how the instructions provided to LLMs hold considerable importance. For example, Chae and Davidson (2024) showed that, provided with a prompt containing detailed instructions, GPT-4o achieves high zero-shot text classification accuracy comparable to GPT-3 Davinci fine-tuned on thousands of labeled examples. This finding suggests that for many scenarios, especially those involving short texts or limited compute resources, prompting LLMs directly can replace costly fine-tuning procedures.

Prompt engineering techniques such as in-context learning (ICL) and chain-of-thought (CoT) prompting have emerged as especially effective. In ICL, the model is provided with task descriptions and demonstration examples shown within the prompt to guide model predictions. CoT prompting further extends this by including intermediate reasoning steps, assisting models in executing complex tasks through explicit step-by-step logic. In the context of annotating educational questions, several studies have shown the benefits of this type of prompt engineering. Xu et al. (2025) introduced inference step generation without relying on example demonstrations, enabling zero-shot chain-of-thought reasoning. Faraby et al. (2024) designed twenty-four prompt variations and demonstrated that, in some contexts, the most effective prompting technique was zero-shot, while few-shot and few-shot + chain-of-thought approaches underperformed. This demonstrates that even with minimal or no training data, zero-shot LLMs can achieve competitive classification results in educational question tagging.

Moreover, Moore et al. (2024) explored different and creative ways to instruct the model to annotate educational questions using a non-label dataset: the simulated expert and simulated textbook approaches, which balanced the need for contextual information across diverse question banks. The simulated expert approach involved prompting the model to generate a discussion among three domain experts to collaboratively decide on the most appropriate concept label. In contrast, the simulated textbook approach asked the model to determine which textbook section the question would most likely appear in, using that as a proxy for the concept. Performance-wise, they reported that their most effective LLM strategy achieved human-evaluated matches of subject-specific concepts for 56% of Chemistry and 35% of E-Learning multiple-choice questions.

Despite the considerable efficacy of LLMs, some limitations remain inherent in automated question annotation tasks. The robustness of LLMs to prompt variations, for example, remains a challenge. Mizrahi et al. (2024) and Xu et al. (2025) observed that even small changes in prompt wording can lead to substantial fluctuations in model performance, making it difficult to generalize across different domains.

Another key challenge lies in aligning the concept labels generated by LLMs with the existing ones in databases. Xu et al. (2025) note that prompts lacking a fixed list of candidate labels may lead the model to generate unexpected or extraneous labels, complicating standardization and evaluation. Therefore, post-processing is necessitated, which often requires addressing LLM output inconsistencies such as multiple answers, irrelevant tokens, or undefined labels (Faraby et al., 2024).

While zero-shot and few-shot prompting strategies using generative models offer strong performance with minimal supervision, they are not always the optimal choice in data-rich environments. In cases where ample annotated data is available, traditional fine-tuning approaches can still surpass zero-shot methods. For example, Faraby et al. (2024) found that fine-tuned encoder-based models such as RoBERTa outperformed zero-shot LLMs in question classification tasks when trained on extensive labeled datasets. This suggests that, despite the flexibility of generative models, BERT-style architectures remain highly competitive when sufficient supervision is available.

Overall, large generative models have demonstrated impressive capabilities in educational question classification, particularly in low-resource settings where annotated data is scarce. However, the success of these models is highly dependent on prompt design, and their outputs often require post-processing to ensure alignment with predefined taxonomies. Moreover, recent findings suggest that in data-rich scenarios, fine-tuned non-generative models, such as BERT, can outperform LLMs, underscoring their continued relevance and strength in educational NLP tasks. This research aims to explore this balance further by evaluating and comparing the performance of both generative and non-generative models in question classification, with a particular focus on practical effectiveness, consistency, and alignment with existing subject-specific concepts.

2.3 Cross-Lingual Application

Another challenge for general tasks using LLMs is the language used in the content. Most large models are primarily trained on English corpora, leading to performance disparities in non-English contexts. As a result, relatively few studies have explored LLM performance in low-resource language settings, especially in the educational field. To address this, researchers have proposed approaches such as cross-lingual prompting (Qin et al., 2023), machine translation, or training on native-language corpora (Isbister et al., 2021; Liu et al., 2024), demonstrating the importance of considering the language when using LLMs. For instance, when dealing with Chinese questions, Xu et al. (2025) applied ZhiPu AI, a model pretrained on extensive Chinese corpora, which demonstrates superior comprehension of questions, highlighting the advantage of language-specific pretraining.

Although it is a challenge, interest in multilingual LLM applications has been growing. Some effective applications of LLMs in multilingual and cross-lingual educational contexts have garnered increasing attention. Focusing specifically on Dutch, Blom and Pereira (2023) applied multilingual BERT-based models to question-answering tasks in Dutch and observed that the multilingual BERT model (mBERT) outperforms Dutch-specific variants like RobBERT and BERTje. Furthermore, the authors highlighted that fine-tuning mBERT on domain-specific data yielded a remarkable improvement, achieving an F1-score of 94.10%, which corresponded to a 226% relative increase compared to the model's performance without fine-tuning. This underscores the importance of both multilingual pretraining and domain adaptation for effective performance in less-resourced languages. In a similar way, Kwak and Pardos (2024) criticized the small amount of research focusing on non-English contexts within educational applications. They presented one of the first empirical findings on addressing disparities in LLM performance across countries and languages within an educational context, achieving great results using GPT 3.5 model and helping to mitigate performance disparities across different languages and educational systems.

To further improve cross-lingual performance, researchers have explored specialized prompting methods. For example, the Cross-Lingual Thought prompting (XLT) approach (H. Huang et al., 2023) explicitly incorporates the target language in prompts, guiding the LLM to adopt the role of an expert fluent in that language for the task at hand. Experimental results demonstrate that XLT not only boosts overall performance on multilingual tasks but also narrows the performance gap between the average and the best-performing languages within each task.

Overall, these studies suggest that multilingual and cross-lingual prompt engineering, along with fine-tuning strategies, are critical to improving fairness and performance in educational AI systems. This is particularly relevant to languages like Dutch and motivates further exploration of multilingual strategies in question classification tasks.

3. Methodology

This section presents the data and methods employed to investigate the effectiveness of LLMs for automatic annotation of assessment questions with subject-specific concepts. Section 3.1 provides a detailed overview of the dataset used in this study and the preparation steps for model training and evaluation. Section 3.2 presents the methods implemented in both generative and non-generative models for the annotation tasks, covering their preprocessing and design decisions. Finally, this section also explains the criteria and metrics used to evaluate and compare the models' performance.

3.1 Data

This study utilizes a labeled dataset in which each question is annotated with a specific topic related to subject concepts. The presence of these labels enables the calculation of metrics such as accuracy, F1-score, and others, thereby facilitating the evaluation of the effectiveness of LLMs in automating the annotation of assessment questions.

The dataset is derived from an open item bank of statistics exercises from various Dutch institutions. It can be found on the website http://www.sharestats.nl, which was developed by a community of instructors from Dutch universities. The bank contains around 5,000 assessment questions in Dutch or English, which were carefully categorized and constructed using a quality-checked procedure (Namesnik-Silvester et al., 2025). All items are shared under a Creative Commons License in a unified format, allowing for easy export to online learning and educational applications. The dataset also includes additional information such as the type of question, language, and level of learning.

During the data cleaning process, some inconsistencies were identified. Nineteen questions were duplicated and were therefore removed. In 26 cases, identical questions were associated with different topics; these were also excluded to avoid confusion when training the models. In addition, questions written in English (17.5%) were removed to focus on a single language and to explore the potential of LLMs in classifying non-English questions. Following these preprocessing steps, the final dataset consisted of 4,028 questions.

To define labels that represent statistical concepts, the developers of the item bank constructed a specific taxonomy. It contains a hierarchical classification structure by topic (see Figure 3.1 for an example of the taxonomy) and was developed to include all statistical topics commonly covered in undergraduate statistics courses for the social and behavioral sciences.

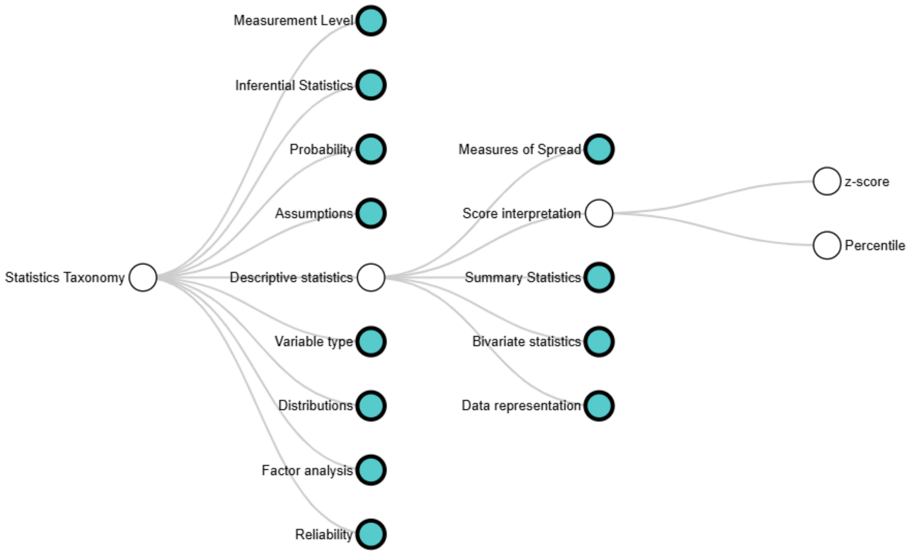

Figure 3.1: Example of the hierarchical taxonomy of statistical concepts.

In total, there are 249 possible categories, written in English, which vary across 9 main levels. For this study, in order to reduce imbalance among the top-level categories, the main levels Measurement Level and Variable Type were combined and assigned to a new category called Type of Variable. Additionally, since the Probability category included only one possible second-level class, the third-level category was used whenever applicable.

3.2 Methods

The main goal of this study, as outlined in RQ1, is to explore the capability of LLMs in annotating assessment questions to support curriculum alignment applications. In particular, it aims to explore both generative and non-generative structures to identify their respective strengths and limitations, addressing both SQ1 and SQ2.

Due to the strong ability to perform zero-shot and few-shot learning, great results in cross-lingual context, and the model's accessibility, this research adopted the generative model from OpenAI, called GPT. Specifically, GPT-4.1 nano ("gpt-4.1-nano-2025-04-14") was used, considering its low costs and strong performance. For comparison, the non-generative multilingual BERT-based model (mBERT) was chosen based on previous studies that demonstrated high performance in Dutch-language contexts (Blom & Pereira, 2023).

By inputting the question text into these models, the goal is to verify their performance in predicting which subject-specific concepts the question aimed to assess. The effectiveness of the models is evaluated by comparing their predictions with the true labels.

All analyses were conducted using Python in the Google Colab environment, chosen for its accessibility and availability of GPU resources. Key libraries used included `pandas` for data manipulation, `scikit-learn` for preprocessing and evaluation, `transformers` for BERT applications, and the `openai` for accessing GPT models.

3.2.1 GPT Application

In the dataset, questions are presented in Markdown format, including LaTeX code for equations where necessary. Considering GPT's ability to understand context and different text formats, `codecs` library was used to convert literal escape sequences (like \n, \t, \\) presented in the text into their actual characters. For instance, instead of \n, it converts into a real line break.

For the prompt design, OpenAI's library allows users to provide instructions using different message roles, called "System" and "User". "System" messages define the model's behavior by specifying high-level instructions, while "user" messages provide inputs and configurations to which the system message instructions are applied. These roles were utilized in the prompt design process to clarify desired behavior and define the annotation task.

In addition, the OpenAI API function allows some optional parameters that help control the output. Two parameters used in this study were `temperature` and `max_tokens`:

- `Temperature`: Controls the randomness or creativity of the model's output. Lower values make the output more deterministic and focused, while higher values increase randomness, allowing for more diverse or creative responses. As this classification task is intended to support automated annotation in future applications, precision and consistency are important considerations. Therefore, a lower temperature value (0.3) was selected.
- `Max_tokens`: Defines the maximum number of tokens in the output. This parameter was useful for controlling cost and output length, and avoiding unnecessarily long completions. As the primary objective is to predict subject-specific concepts, a maximum of 150 tokens was deemed sufficient.

To answer the first subquestion (SQ1), which aims to explore different prompt formats and provided contextual information, the following strategy was employed. First, a simple and straightforward **Baseline** prompt was designed to directly ask the model to classify a given question according to statistical concepts. This prompt serves as a reference point, allowing for comparisons with more complex prompting strategies. In addition, to ensure comparability with mBERT and to evaluate exact matches with the labeled dataset, the model was restricted to selecting from a predefined list of topics. Finally, an instruction was added requiring the model to provide a detailed explanation alongside its answer, allowing assessment of the reasoning behind its decisions.

The **Baseline** prompt is illustrated in Figure 3.2:

Baseline Prompt

System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts.

User: Analyze the following question to determine the most appropriate topic from this list of statistical concepts:
`[‘assumptions’, ‘factor analysis’, ‘descriptive statistics’, ‘inferential statistics’, ‘distributions’, ‘reliability’, ‘type of variable’, ‘probability’].`
Your task is to examine the question text below and select the most fitting topic based on the knowledge components and skills the question assesses.
Question: `{input}`
Please provide both the topic and rationale as follows:
Final Topic: `<exact topic from the list above>`
Rationale: `<detailed explanation for the selected topic>`

Figure 3.2: Baseline prompt used for annotation task with fixed-label list.

To initialize prompt engineering experiments, a cross-lingual prompting strategy was tested to explore the multilingual context of the task. Inspired by the work of H. Huang et al. (2023), this approach provides additional information by explicitly guiding the model to assume the role of an expert in interpreting questions written in Dutch. To compare the results with the actual label and to clarify the task, the following instruction was added: “Important: Although the question is in Dutch, your output must be in English.” (see Table A.1 in the appendix for more details).

Subsequently, the variations of in-context learning (ICL) and chain-of-thought (CoT) prompting were tested. These techniques aim to enhance the model’s performance by guiding its reasoning process:

- One-shot learning: Providing an example of a question and the expected answer.
- Few-shot learning: Providing three examples of questions with their corresponding expected answers.
- Chain-of-thoughts (when applicable): Including a step-by-step explanation of the reasoning process the model should follow to classify the question.

The examples used in these prompting strategies were selected through a careful manual analysis of the questions that the model had correctly classified in the earlier tested approaches. The goal was to choose examples that were simple and clear, ensuring they could effectively guide the model’s reasoning in the ICL settings (see Table A.1 in the appendix for more details).

Besides these prompt format techniques, to evaluate the impact of contextual information, such as the fact that the questions were applied in university-level statistics courses, and to vary prompting instructions, two additional strategies were developed based on the work of Moore et al. (2024). They were defined as follows:

- 1. **Experts prompt:** Uses a tree-of-thought approach, directing the LLM to simulate a discussion among three expert instructors to determine the appropriate topic. After this simulated discussion, the experts were expected to choose the main topic targeted by the question.
- 2. **Textbook prompt:** Instructs the LLM to identify the specific topics that would be covered on a textbook page if the given question were included.

The examples are shown in Table A.2 in the appendix section.

To finalize, while the previous prompting strategies rely on a fixed-label list to guide the classification task, an alternative approach was explored to assess the model's ability to generate relevant subject concepts more freely. This generative prediction classification strategy allowed the model to produce labels without explicit constraints, enabling the exploration of its capacity to infer and articulate concepts beyond fixed categories, which is particularly useful in scenarios where providing contextual information may not be feasible. The same prompt engineering strategy described previously was applied: **Cross-Lingual**, **One/Few-shot**, **CoT**, **Experts**, and **Textbook** approaches (examples can be found in Table A.3).

For the evaluation and comparison of all prompts applied, although a temperature of 0.3 was used to make the model more deterministic, the **Baseline** prompt was executed five times to explore the potential variability in the output. The standard deviations of performance scores were relatively low (around 0.7% for prompts with a predefined list of labels and 0.5% for the generative classification approach). Due to these small variations and the cost of generating multiple completions for each prompt, the variability observed in the **Baseline** prompt was used as a proxy for the others. This approach allowed for the evaluation of whether improvements were attributable to the prompt configuration or merely the result of natural variation inherent to generative models. Future work could incorporate larger samples to better quantify variability.

3.2.2 MBert Application

To answer the second subquestion (SQ2), which aims to compare two different types of LLMs in annotating questions, the non-generative multilingual BERT-based model (mBERT) was also considered for this analysis. BERT (Bidirectional Encoder Representations from Transformers) models are pretrained language representations designed to capture contextual information from both the left and right context in text. The multilingual variant, mBERT, is trained on Wikipedia text from 104 languages, enabling it to effectively handle multiple languages, including Dutch. Previous studies have demonstrated mBERT's strong performance on Dutch-language NLP tasks (Blom & Pereira, 2023), motivating its use here for comparison against generative models.

An advantage of BERT-based models is their ability to be fine-tuned for specific downstream tasks, allowing them to adapt more effectively to the requirements of the project. To enable this, the dataset was split using a train-validation-test approach: 80% of the data was allocated to training and validation, with 20% of that portion reserved for validation. The remaining 20% of the full dataset was set aside for testing.

The preprocessing of the question text differed from that applied in the GPT model (see Section 3.2.1). Markdown characters were cleaned to prevent tokenization issues and reduce noise during training, thereby improving the model's ability to learn from the actual content of the questions. LaTeX characters were transformed into text representation using the `pylatexenc.latex2text` library. In addition, some questions contained references to downloadable files or images that were not included in the dataset, such as `[SPSS file] (eur-descriptive-251-en-data.sav)` or `question.png`. In these cases, image file references (e.g., `.jpg`, `.png`, `.jpeg`) were replaced with the word "image" and download links were simplified by removing the file path while keeping the label (e.g., `[SPSS file]`). To finalize, some Dutch stopwords were also removed using the `nltk` library.

Considering that this is a multi-class classification task, the macro-averaged F1 score was chosen as the target metric to be maximized during the training process. This metric treats each class equally, regardless of how many examples each class has. The F1 score is defined as:

$$\mathrm{F1}_i = \frac{1}{N}\sum_{i=1}^{N}\frac{2\cdot \mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

where:

- N is the number of classes,
- $\mathrm{Precision}_i$ and Recall_i are the precision and recall for class i,
- $\mathrm{F1}_i$ is the F1 score for class i.

Therefore, the macro-averaged F1 is defined by the average of each class F1 score:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{F1}_i \ ,$$

While training the model, the `optuna` library was used to explore different combinations of hyperparameters within a specified search space. The specified ranges included a learning rate between 1e-5 and 5e-5, a number of epochs from 3 to 10, and training batch sizes of 8, 16, or 32. By using a Bayesian optimization, the hyperparameter search is initiated with a random combination of parameters from those ranges. Based on the performance of previous trials, subsequent configurations are chosen strategically to improve efficiency. To finalize, it outputs the best combination of parameters that maximizes the performance.

Due to the presence of class imbalance in the dataset, a custom focal loss function was employed to adjust the model's attention to underrepresented classes during training. A class weight was computed using `scikit-learn`'s `compute_class_weight` function, which assigns higher importance to minority classes based on their inverse frequency. Subsequently, these weights were incorporated into a modified focal loss function, which replaced the default cross-entropy used by the Hugging Face library `Trainer` function. This approach helped prevent the model from being biased toward the majority classes and ensured more balanced learning across categories.

3.2.3 Hierarchical Model

Due to the hierarchical structure of the label concepts variable and based on the Local Classifier Per Parent Node approach described by Silla and Freitas (2011), the following analyses were conducted:

1. **Main-level classification:** Analyze the models' performance in predicting only the main level of the taxonomy. For models such as BERT, limiting the classification task to just eight first-level classes (as opposed to 249 possible categories) effectively reduces the complexity of the task. This reduction also mitigates issues related to data imbalance, potentially improving overall model performance.

2. **Second-level classification:** For each predicted main-level class, analyze the model's ability to predict the corresponding second-level of the taxonomy. Deeper taxonomy levels provide more detailed information, improving interpretability and utility of the classification in educational contexts. Due to sample size and class imbalance, only four of the eight main-level categories were included in the second-level classification analysis. The excluded categories contained second-level labels with fewer than three examples each, which were insufficient to support reliable training. Moreover, such limited data leads to unstable evaluation metrics, where the classification outcome of a single sample can disproportionately influence accuracy and macro F1 scores, thereby compromising the reliability of performance assessments.

   Both classification methods were applied in mBERT and GPT models for comparison. Other levels of the taxonomy were not considered due to sample size.

During preprocessing, it was observed that some questions were associated with multiple labels. The following cleaning procedures were applied. For main-level classification, questions annotated with more than one distinct main-level label were removed, representing 139 questions in total (3%). For second-level classification, questions that shared the same main-level label but differed in second-level labels were also removed (132 in total, representing 3%) to avoid ambiguity. Finally, questions annotated with multiple concepts that shared the same main and second-level labels were retained.

3.2.4 Evaluation

To ensure a fair and reliable comparison between mBERT and GPT models, all performance metrics were calculated using the same test dataset employed in the mBERT application. This approach attributes any observed performance differences to the models themselves rather than to variations in the underlying data.

To evaluate their performance, accuracy was selected as one of the metrics, measuring the proportion of predictions that exactly match the true labels. For GPT application, it is important to note that even when prompts included a predefined list of topics, some predictions deviated from this list, requiring text normalization or manual adjustments to align with the ground truth labels. For instance, the model might generate a descriptive phrase such as “interpretation of regression equations (related to ‘inferential statistics’)” instead of the exact label. In such cases, the output was manually mapped to the standardized label “Inferential Statistics.” Other typical variations included plural forms (e.g., “probabilities” mapped to “Probability”) or synonymous expressions (e.g., “factorial analysis” mapped to “Factor Analysis”).

Evaluating the performance of LLMs in question annotation also requires the use of multiple complementary metrics. Accuracy provides a straightforward indicator of performance; however, in scenarios with imbalanced class distributions, such as in this study, it can be misleading, as it tends to be biased toward the majority classes. To address this issue, the macro-average F1-score was used as a complementary metric. It equally weights precision (the proportion of relevant predicted labels) and recall (the proportion of relevant actual labels) across all classes, ensuring that performance on minority classes is fairly represented.

In contrast, evaluating prompts that did not include a predefined list of labels required a different strategy. In this scenario, the model was free to generate topics without being constrained by a fixed label set. To handle this, two main evaluation approaches were employed:

1. **Semantic similarity score:** Measuring a semantic similarity between the true

topic labels and the model's predicted topics. Rather than relying on exact string matches, the similarity score captures how close the predicted topic is to the true topic, allowing some flexibility for variations in wording. Sentence embeddings for each label were generated using the `SentenceTransformer` model `all-MiniLM-L6-v2`, which maps text into a dense vector space where semantic similarity can be computed. Cosine similarity was then calculated between each pair of predicted and true label embeddings. A threshold of 0.5 was set to determine whether a prediction was considered correct; if the similarity exceeded this threshold, the prediction was counted as accurate. To ensure fair comparison, each predicted topic was evaluated against all labels in the hierarchical taxonomy, allowing the model to match on any level of granularity within the topic structure.

2. **LLM evaluation:** Using a structured prompt designed for GPT-4.1-nano to determine whether a predicted topic label is semantically equivalent to the true label (represented by the whole hierarchical taxonomy). The prompt presents both the true and predicted labels and asks the model to judge whether the prediction is a valid match based on conceptual equivalence, regardless of differences in wording or hierarchy. An example of the prompt used is shown in Figure 3.3. The final evaluation metric was calculated as the proportion of matched questions overall.

GPT Judgment Evaluation Prompt

System: You are an expert in educational content evaluation, specializing in university-level Statistics. Your task is to determine whether a predicted statistical topic is semantically close enough to a ground truth label used in a curriculum classification system.

User: You will be given two pieces of information:
1. A ground truth label that represents the topic of a statistics question, taken from a course outline. This label may include hierarchical structure (e.g., "Probability / Elementary Probability / General Rules / Addition rule").
2. A predicted label generated by a language model, based on the content of the same question.
Please determine whether the predicted label is a valid match for the true label. Consider the semantic similarity of the core statistical concepts and not the formatting. Focus on whether the prediction refers to the **same statistical topic or idea**, even if expressed differently.
Respond with:
Match: Yes or No
Explanation: A brief justification of your decision.
Ground Truth Label: {true_label}
Predicted Label: {prediction}

Figure 3.3: GPT judgment evaluation prompt for comparing predicted and true statistical topic labels.

In addition, to provide a more comprehensive evaluation of the generative model's performance, all prompts included an instruction for the model to produce a detailed explanation along with its prediction. A manual evaluation of a subset of these explanations was conducted to gain qualitative insights into the model's behavior, including its strengths, interpretability, and common failure modes.

Overall, a combination of quantitative (macro-averaged F1 and accuracy) and qualitative (manual evaluation of GPT explanations) evaluation strategies provides a comprehensive picture of LLM effectiveness in annotating educational questions.

4. Results

This section presents the results of the study. First, an exploratory analysis was conducted to gain a better understanding of the types of questions present in the dataset. Next, Section 4.2 investigates the impact of prompt engineering on the performance of generative models, aiming to identify strategies that yield optimal results. Finally, Section 4.3 presents a comparison between the best-performing prompt and results of a non-generative model to understand in which scenarios each approach performs better, offering insights into their suitability for curriculum alignment tasks.

4.1 Exploratory Analysis

To better understand the structure and content of the dataset, an exploratory analysis of the questions was conducted. The results show that the majority (91%) are multiple-choice questions with four answer options. Additionally, 82% of the questions target learning objectives related to statistical literacy. These questions mainly represent three task categories: conceptual comprehension, procedural calculation, and interpretation of statistical outputs.

Table 4.1 presents the distribution of questions across the main levels of statistical concepts according to the taxonomy used in the dataset. The majority of questions (62.3%) fall under Inferential Statistics, making it the most dominant category with 3,044 entries. This is followed by Descriptive Statistics (15.4%) and Probability (10.1%). The remaining categories account for less than 5% of the total. Overall, this distribution reveals a presence of class imbalance, which may impact the performance of classification models if not properly addressed during training. This issue was appropriately mitigated, as discussed in Section 3.2.2.

Main Level	**Frequency**	**Relative Frequency**
Inferential Statistics	3044	62.3%
Descriptive Statistics	755	15.4%
Probability	492	10.1%
Distributions	230	4.7%
Assumptions	132	2.7%
Reliability	104	2.1%
Factor Analysis	66	1.4%
Type of Variable	64	1.3%

Table 4.1: Frequencies and relative frequencies of main-level topics in the dataset.

To explore the second-level classes, the number of subcategories within each main level was evaluated. This is presented in Table 4.2. It is noted that the number of second-level classes differs across the main categories, indicating varying degrees of complexity or granularity within each topic. For example, Descriptive Statistics, Distributions, and Reliability have the fewest subcategories, suggesting these topics are relatively straightforward or less diverse. In contrast, Factor analysis and Inferential Statistics stand out with the highest number of subcategories, 13 and 11, respectively, highlighting their broad nature. Overall, the varying levels of detail across topics emphasize that some statistical concepts are much more granular than others, which might affect the performance of the models.

Main Level	**N° of Second-Level Classes**
Factor Analysis	13
Inferential Statistics	11
Assumptions	7
Probability	7
Type of Variable	7
Descriptive Statistics	4
Distributions	4
Reliability	3

Table 4.2: Number of second-level classes within each main-level topic.

4.2 Impact of Different Prompt Strategies

This section addresses the first subquestion of this study (SQ1), which aims to evaluate how different prompt formats and the inclusion of curricular context influence the performance of generative models. It aims to identify which strategies yield optimal results.

4.2.1 In-Context Learning Prompts

To assess the impact of prompt format on annotation performance, different prompting techniques were investigated. Starting with a Baseline prompt (see Figure 3.2), successive enhancements of different techniques were used and evaluated, aimed at improving classification accuracy and macro F1 score. All prompts were provided with a predefined list of labels, instructing the model to select the most appropriate one for each question. The results are presented in Figure 4.1.

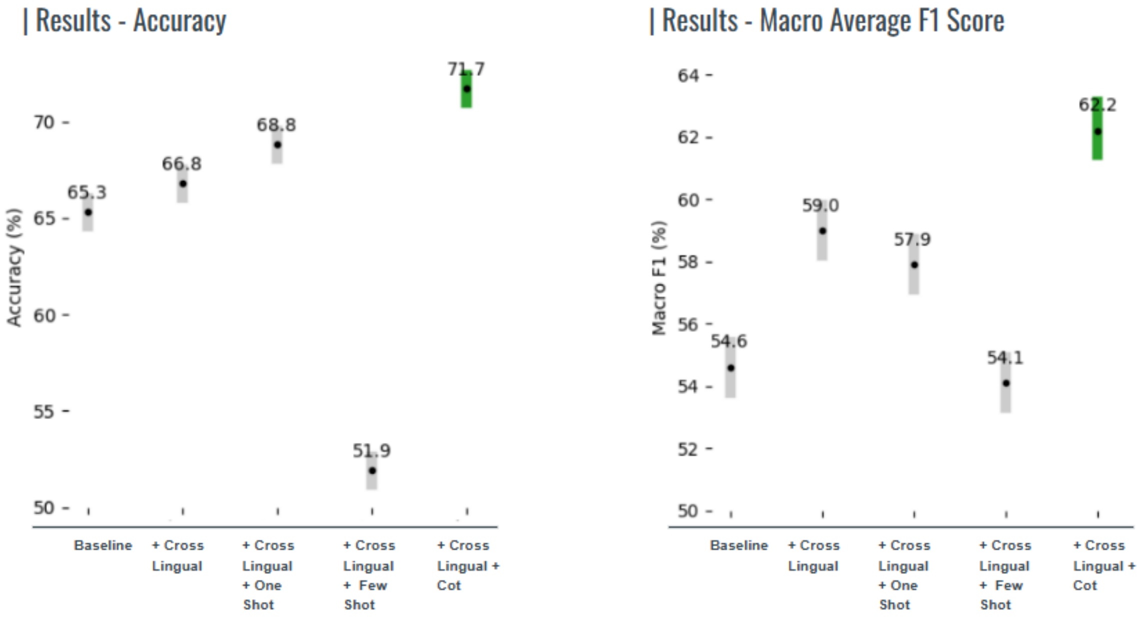

Figure 4.1: Comparison of In-Context Learning settings. Boxes reflect a proxy for variability of the model estimated based on five runs of the Baseline prompt with identical inputs.

The **Baseline** achieved an accuracy of 65.3% and a macro-averaged F1 of 54.6%, serving as a reference point. Adding a cross-lingual component that emphasizes the use of Dutch resulted in improvements for both metrics, increasing accuracy to 66.8% and macro F1-score to 59.0%. The results suggest an increase in the model's capacity to classify difficult classes and improve performance across several categories.

Following this, different in-context learning (ICL) scenarios were tested: **One-shot** learning, **Few-shot** learning, and **Chain-of-Thought** prompting (CoT). **One-shot** learning, which included an example question labeled as Descriptive Statistics, improved the model's accuracy from 66.8% to 68.8%. However, it did not result in an overall increase in macro F1, which decreased slightly from 59.0% to 57.9%. Analyzing class-specific F1 scores shows that while the model's performance improved notably for the example class, Descriptive Statistics, rising from 56.8% to 64.6%, it declined for other topics, resulting in an overall decrease in macro F1 (class-specific F1 scores are reported in Table A.4 in the appendix section).

To encourage broader generalization beyond the Descriptive Statistics class, a **Few-shot** technique including examples from two other classes (Assumptions and Inferential Statistics) was introduced. However, this resulted in a notable performance drop: accuracy fell by 14.9 percentage points, from 66.8% to 51.9%, and macro F1 decreased by 4.9 percentage points, from 59.0% to 54.1%. While the F1 score for Descriptive Statistics improved from 57% to 66%, the two additional example classes, Inferential Statistics and Assumptions, saw declines from 79.8% to 60.8% and from 60.7%

to 19.6%, respectively. The results suggest that increasing the number of examples introduced additional complexity, which may have confused the model rather than improving its generalization.

In all prompting strategies, GPT was instructed to provide a brief explanation justifying its label choice for each question. When manually evaluating some examples where GPT provides a justification of its label choice for each question, a clear illustration of the few-shot method's limitations is found in the following question:

> **Question:** "[...] a researcher performed a multi-way ANOVA for 'knowledge of shape' with the factors 'residential area' and 'sex'. Use the average diagram shown to evaluate the following two statements. Assume that all differences you see are significant [...]"

All other prompting strategies correctly classified this question under Inferential Statistics, with rationales such as:

> "The question involves interpreting the results of a multi-way ANOVA, which is a statistical inferential technique."

However, the **Few-shot** prompt misclassified it as Assumptions. This confusion likely arose from the phrase, *"Assume that all differences you see are significant,"* which is a standard instruction unrelated to the statistical Assumptions category. This suggests that the model was overly influenced by the examples provided during training, one of which focused on Assumptions, highlighting the method's sensitivity to example selection.

Finally, providing the model with step-by-step reasoning instructions through **CoT** prompting led to a meaningful improvement over all previous prompting strategies, achieving 71.7% accuracy and 62.2% macro F1. These results reveal a distinct gap compared to the **Baseline** with the cross-lingual prompt in both metrics. Given that the proxy for variability across runs was below 1%, this difference indicates a consistently better performance likely attributable to the CoT prompting strategy rather than random variation. Additional evidence comes from the improvements in F1 scores across all individual classes.

Manual evaluation of some outputs revealed that, while **Baseline** prompt often misclassified questions by fixating on isolated keywords rather than understanding the broader conceptual context, CoT encouraged more structured and thoughtful reasoning. For instance, a question containing terms such as *"2x2 mixed factorial"* and *"2x2 repeated measures factorial"* was incorrectly labeled as Factor Analysis by all other prompts, likely influenced by the term "factorial" without fully considering the question's context. In contrast, the **CoT** prompt correctly labeled it as Inferential Statistics and supported this decision with the following explanation:

> *"The focus is on understanding how these factors influence perceptions,*

which implies analyzing differences between groups or conditions and making inferences about the population based on the sample data. This aligns with the concept of inferential statistics, which involves drawing conclusions about populations from sample data, often through hypothesis testing or analysis of variance (ANOVA)."

Highlighting that **CoT** prompting enables the model to reason more thoroughly and avoid superficial keyword associations, resulting in more accurate topic identification.

4.2.2 Context-Enriched Prompting Strategies

Given the previous results, two additional methods (described in Section 3.2.1) were tested to enhance the **Baseline** prompt by incorporating contextual information about the background of questions and exploring alternative approaches to question annotation. The **Experts** prompt guided the model to simulate a discussion among instructors to collaboratively decide on the appropriate topic, while the **Textbook** prompt instructed the model to identify topics that would appear on a textbook page covering the question. The detailed results are shown in Figure 4.2.

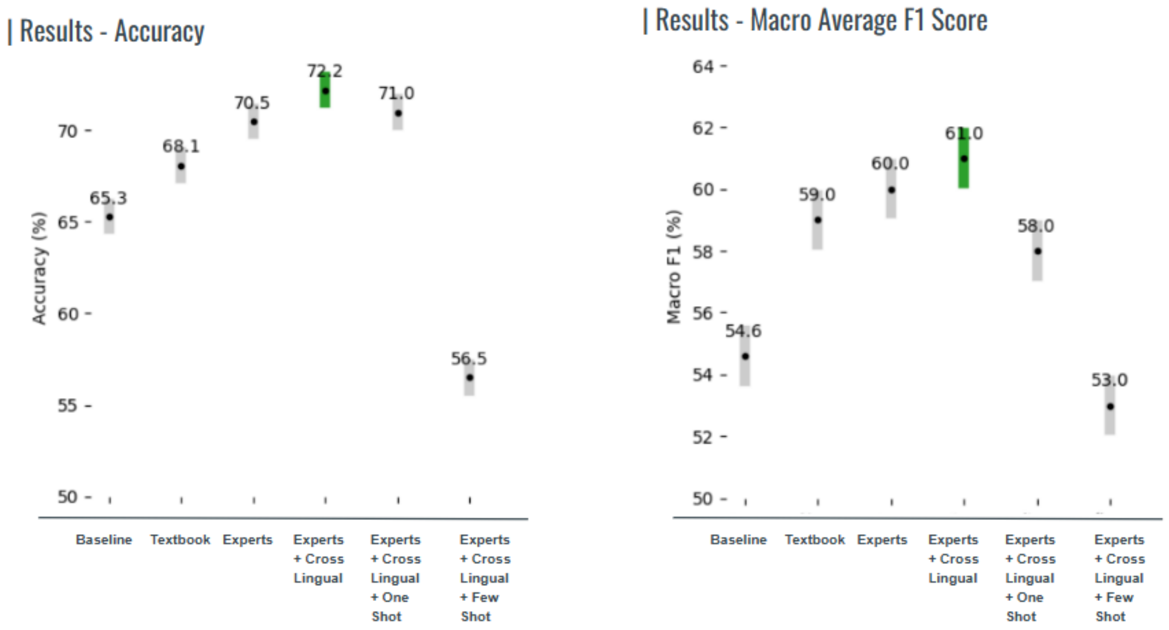

Figure 4.2: Comparison of different prompt strategies including Experts and Textbook. Boxes reflect a proxy for variability of the model estimated based on five runs of the Baseline prompt with identical inputs.

The initial comparison includes the **Baseline** and two prompts: **Experts** and **Textbook**. The best-performing strategy, **Experts**, was further extended with **Cross-Lingual** and **Few/One-shot** variants. Both the **Textbook** and **Experts** methods outperformed the **Baseline** in terms of accuracy and macro F1. The **Experts**' approach, in particular, achieved the highest overall performance with 70.5% accuracy and a 60.0% macro-averaged F1 score. This may be related to the provided structure instructions closely resembling chain-of-thought prompting, a method shown to facilitate step-by-step reasoning and previously demonstrated in Figure 4.1 to improve model outcomes.

When evaluating some examples of **Experts** prompt outputs, the similarity with **CoT** prompt becomes evident. Considering the following multiple-choice question:

> **Question:** “Which word is missing? The ‘proportion of explained variance’ tells something about the:
> a. effect size;
> b. reliability;
> c. significance;
> d. validity.”

The Baseline model misclassified the question under Reliability, reasoning that:

> *“In the context of the answer options, ‘reliability’ relates to the consistency and dependability of a measurement instrument, which is directly connected to the proportion of variance explained by the measurement.”*

This interpretation narrowly focuses on the keyword “reliability,” overlooking the broader inferential context of the question. In contrast, the **Experts** prompt correctly identified the question as belonging to Inferential Statistics, justifying:

> *“This concept is fundamentally linked to the assessment of how well a statistical model explains the variability in the data, which is a core aspect of inferential statistics.”*

This example underscores how structured prompting techniques help the model interpret questions more contextually, leading to more accurate predictions.

Since the **Experts** method already employed a CoT style, no additional step-by-step prompting was tested. Instead, **One-shot** and **Few-shot** approaches were evaluated to see if they could further improve performance. In line with previous observations, the **One-shot** approach yielded less balanced results, as reflected by a lower macro F1 score. In addition, the **Few-shot** method produced the poorest overall performance, with an accuracy of 56.5% and a macro-averaged F1 of 53.0%.

4.2.3 Summary of Prompting Strategies

To conclude the analysis of prompt engineering strategies, this subsection presents a detailed comparison of the best-performing prompts across individual classification categories. Table 4.3 reports the F1 scores by class for the **Baseline** prompt, the most effective in-context learning setup (**Baseline + Cross-Lingual + Chain-of-Thought**), and the best-performing context-enriched approach (**Experts + Cross-Lingual**). This comparison highlights which model performs best for each individual class.

Class	Baseline	Baseline + Cross Lingual + CoT	Experts + Cross Lingual	Number of Samples
assumptions	50.8%	60.0%	65.4%	19
descriptive statistics	58.1%	57.5%	57.1%	119
distributions	22.0%	25.2%	24.6%	42
factor analysis	42.9%	66.7%	57.1%	8
inferential statistics	78.6%	83.5%	83.2%	483
probability	68.0%	71.3%	71.4%	95
reliability	50.0%	66.7%	60.0%	10
type of variable	66.7%	66.7%	66.7%	8
macro avg	54.6%	62.2%	60.7%	784

Table 4.3: F1 score by class and number of aamples for the **Baseline** prompt compared to the best-performing strategies (**Baseline + Cross-Lingual + CoT** and **Experts + Cross-Lingual**). Results highlighted in blue indicate the best-performing setting for each category.

The results show that **Baseline** prompt performance varies considerably across different classification categories. Certain topics, such as Distributions (22%) and Factor Analysis (42.9%), proved more challenging for the model to classify, whereas other categories, such as Inferential Statistics (78.6%) and Probability (68.0%), achieved much higher F1 scores. These results indicate that the model's ability to correctly classify questions can also be dependent on the specific conceptual category being tested. Despite these differences, both enhanced strategies (**Baseline + Cross-Lingual + CoT** and **Experts + Cross-Lingual**) outperform the **Baseline** across nearly all classes, reinforcing that prompt engineering can substantially improve performance, particularly highlighting the positive impact of incorporating step-by-step reasoning and multilingual contextual information.

When comparing the performance of **Baseline + Cross-Lingual + CoT** with the **Experts + Cross-Lingual**, the first prompt achieves higher F1 scores across more classes, which is also reflected in its superior macro F1 score. However, it is important to emphasize that the overall difference in performance is relatively small and could be attributed to the inherent variability in the model's outputs, which was estimated at around 0.7% across runs. Therefore, both prompting strategies should be considered valid and effective approaches for the assessment question annotation task.

Due to its consistently stronger performance across a broader range of classes and slightly higher macro F1 score, **Baseline + Cross-Lingual + CoT** was selected as the representative prompt for further comparison with mBERT-based models in Section 4.3.

4.3 Comparing generative and non-generative models

To answer the second subquestion (SQ2) of this study, which aims to explore different models in annotating questions, a comparison of the performance of GPT-4.1-nano against a non-generative model (mBERT) was employed. Given the hierarchical structure of the dataset's labels, the annotation task was carried out in two stages. The first stage involved predicting the main-level taxonomy categories, consisting of eight distinct classes. In the second stage, using only the four main-level classes with sufficient data, a secondary classification was performed to predict the corresponding second-level categories. As a result, separate models were trained for each level, allowing for evaluation using both accuracy and macro-averaged F1 score.

Unlike GPT models, which can be prompted flexibly, mBERT is restricted to predicting among predefined labels. Therefore, for a fair comparison, the results are compared with the best-performing prompt of GPT-4.1-nano that used constraints to select from the same predefined label sets, the **Baseline + Cross Lingual + CoT** approach.

When training mBERT on the annotated statistics questions, the model used the macro-averaged F1 score to maximize its performance. By trying many settings, the best-performing model for the main-level classification used the following hyperparameters: `learning_rate` = 4.87e-05, `num_train_epochs` = 5, and `per_device _train_batchs` = 8. Presented in Table 4.4, when tested in the 20% dataset (784 questions), mBERT achieved 91.7% accuracy and a macro-averaged F1 score of 83.7% on the test set. Although GPT-4.1-nano achieved competitive results without task-specific fine-tuning, mBERT outperformed it when trained on the classification task.

Model	**Accuracy**	**Macro F1 Score**
GPT	71.7%	62.2%
mBERT	91.7%	83.7%

Table 4.4: Performance comparison between GPT-4.1-nano and mBERT on main-level classification. Results highlighted in blue indicate the best-performing model.

Table 4.5 shows the F1 score achieved by each model for individual classes. These results highlight mBERT's consistent performance across categories, even those with fewer training samples, reinforcing its ability to handle class imbalance more effectively than GPT-4.1-nano.

Main-Level	GPT	mBERT	Number of Samples
assumptions	60.0%	81.3%	19
descriptive statistics	57.5%	86.2%	119
distributions	25.2%	76.2%	42
factor analysis	66.7%	87.5%	8
inferential statistics	83.5%	95.6%	483
probability	71.3%	90.3%	95
reliability	66.7%	66.7%	10
type of variable	66.7%	85.7%	8

Table 4.5: F1-score (%) by Main-Level and Number of Samples for GPT-4.1-nano and mBERT. Results highlighted in blue indicate the best-performing model.

Subsequently, both models, GPT and mBERT, were evaluated on their ability to predict the second-level category. Due to small samples, only four of the main eight classes were used. Training mBERT with diverse settings, the optimal hyperparameters for the classifiers varied by main-level categories:

- **Inferential Statistics**: `learning_rate` = 1.06e-05, `num_train_epochs` = 6, `batch_size` = 8
- **Descriptive Statistics**: `learning_rate` = 2.10e-05, `num_train_epochs` = 5, `batch_size` = 16
- **Probability**: `learning_rate` = 3.52e-05, `num_train_epochs` = 8, `batch_size` = 8
- **Distribution**: `learning_rate` = 2.77e-05, `num_train_epochs` = 8, `batch_size` = 8

To compare the results of these models with the GPT results, the same best-performing prompt (**Baseline + Cross Lingual + CoT**) was used to predict the second-level by only replacing the predefined label list with the possible topics from each category. While the prompt demonstrated strong performance in the main-level classification task, its effectiveness did not carry over to the second-level classification. As shown in Table 4.6, the accuracy and macro-averaged F1 scores for GPT-4.1-nano have lower results across all subcategories.

Comparing the models, the fine-tuned mBERT model consistently outperformed GPT -4.1-nano across all second-level tasks, with accuracy improvements of approximately 20 to 47 percentage points and macro F1 gains between 15 and 37 points. Notably, mBERT performed particularly well in Descriptive Statistics and Probability, where it showed the largest gains over GPT.

Overall, these findings demonstrate the clear advantage of task-specific fine-tuning for hierarchical question classification, as well as underscoring the importance of adapting generative models to domain-specific tasks to achieve higher accuracy and more reliable performance.

Main Level	Model	Accuracy	Macro F1
Inferential Statistics	GPT	57.2%	47.5%
	mBERT	79.9%	73.2%
Descriptive Statistics	GPT	63.3%	53.8%
	mBERT	90.8%	87.1%
Probability	GPT	30.3%	33.4%
	mBERT	77.6%	70.8%
Distribution	GPT	45.2%	43.3%
	mBERT	88.1%	59.2%

Table 4.6: Second-level classification results showed by Main Level for GPT-4.1-nano and mBERT. Results highlighted in blue indicate the best-performing setting for each metric.

4.4 Generative Prediction Classification

The generative prediction classification approach, applied specifically to GPT-based models, evaluates the model's ability to produce free-form text responses that represent the subject-specific concepts assessed by each question. Unlike previous classification methods applied, which require selecting from a fixed set of predefined labels, this no-label configuration challenges the model to produce meaningful labels for the questions without explicit label options. To assess the alignment between the model's free-form responses and the true labels, cosine similarity and GPT-based judgment were used.

It is important to note that, while the variability estimated captures potential fluctuations in the model's output generation, the GPT-based evaluation itself may also introduce additional variability. However, due to computational limitations, the evaluation variability was not empirically measured in this study and should be considered a potential source of uncertainty when interpreting the results.

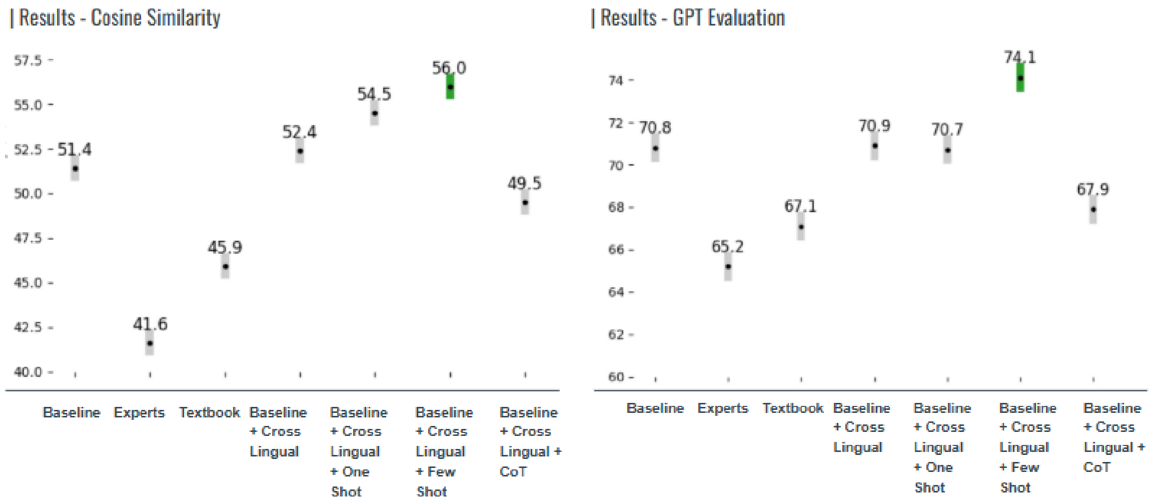

Figure 4.3: Comparison of prompt strategies without a predefined list. Boxes reflect a proxy for variability of the model estimated based on five runs of the Baseline prompt with identical inputs.

Overall, some discrepancies between cosine similarity and GPT-based evaluation were observed, largely due to differences in phrasing and label specificity. Manual inspection showed that while the original labels were often detailed and hierarchical, the model typically generated more concise or paraphrased descriptions. As cosine similarity operates on surface-level embeddings, it may penalize outputs that are semantically correct but differ in expression. For instance, the label "Descriptive Statistics / Summary Statistics / Measures of Spread / Variance" was predicted as "sample variance calculation", resulting in a cosine similarity of just 39%, which was below the evaluation threshold. Nevertheless, GPT-based judgment deemed the output accurate, recognizing that the core meaning was preserved.

The initial comparison includes the **Baseline** and two prompts adapted from Moore et al. (2024): **Experts** and **Textbook**. The best-performing strategy, **Baseline**, was further extended with **Cross-Lingual**, **Few/One-shot**, and **Chain-of-Thought** variants. The **Baseline** prompt, which directly instructs the model to define an appropriate label for the question (see Table A.3 for more detail), achieved a cosine similarity above 50% in 51.4% of the questions. Additionally, 70.8% of the questions were evaluated by GPT as a semantic match with the original label.

In contrast to previous results, both the **Experts** and **Textbook** prompts underperformed. The **Experts** prompt yielded only 41.6% of cases above the cosine similarity of 50% and 65.2% agreement in GPT-based evaluation, while the **Textbook** prompt performed slightly better, with 45.9% and 67.1%, respectively. These findings suggest that, for the open-ended generation task, the more direct instruction provided by the **Baseline** prompt resulted in more accurate and semantically aligned outputs.

Analyzing some of the outputs of both of these methods it is noticeable that the

Textbook produced overly broad labels resembling section headers rather than specific concepts. For instance, a question labeled as Factor Analysis was predicted as *"data reduction and reliability assessment"*. Although related, the GPT-based evaluation concluded it was not an accurate match, highlighting that *"The predicted label describes broader concepts that may include factor analysis but also encompass other methods"*. This illustrates how such excessive scope can reduce alignment with specific educational objectives.

On the other hand, some examples of **Experts**'s outputs showed that it frequently produced responses framed in terms of reasoning or procedural actions. Although semantically related, these outputs were often abstract or overly generalized. For example, the label *"Descriptive Statistics / Score Interpretation / z-score"* was predicted as *"calculating probabilities and percentages for normally distributed data"*.

Building on the **Baseline** prompt, additional strategies were explored, including **Cross-Lingual**, **CoT**, and **One/Few-shot** prompts. The **Cross-Lingual** variant slightly improved performance, with 52.4% achieving a cosine similarity higher than 50% and 70.9% evaluated as a semantic match. In contrast to earlier findings from the classification task, the inclusion of examples further boosted performance: the **One-shot** prompt produced 54.5% high-similarity outputs and achieved a 70.7% match rate according to GPT evaluation. The **Few-shot** setting yielded the best overall results, with 56.0% of responses exceeding the similarity threshold and 74.1% rated as accurate semantic matches. When analyzing some of the outputs per main-level categories, the **Few-shot** learning approach demonstrated a notable improvement in semantic match evaluated from GPT, particularly in categories explicitly represented in the example set. For instance, for questions labeled as Assumptions, the **Few-shot** results achieved 94% semantic match based on GPT, compared to 42% for the **Baseline** results (see Table A.6 for more details).

A concrete example illustrates the benefit of **Few-shot** prompting. When presented with a question focused on interpreting the results of a Levene's statistical test in the context of covariance analysis, the **Baseline** prompt predicted:

> *"Final Topic: Covariance Analysis (ANCOVA)*
> *Rationale: The mention of performing a covariance analysis and the inclusion of an output related to the Levene's test (which assesses homogeneity of variances) indicates that the focus is on an analysis that adjusts for covariates to compare group means. The question assesses understanding of when and how to interpret the results of an ANCOVA."*

In contrast, the Few-shot prompt correctly shifted focus toward the conceptual purpose of the test, aligning more closely with the intended label:

> *"Final Topic: Assumptions*

> *Rationale: This question involves interpreting the results of a Levene's test, which is used to assess the homogeneity of variances—a key assumption in covariance analysis (ANCOVA). It directly relates to the assumptions underlying the statistical method."*

This improvement is particularly noteworthy because Assumptions was one of the concepts included in the example set, underscoring the model's capacity to generalize from well-annotated instances.

Finally, adding step-by-step reasoning in the **CoT** prompt led to a slight drop in performance, with only 49.5% of outputs reaching high cosine similarity and 67.9% meeting GPT's semantic matching criteria.

5. Discussion

This study investigated the capabilities of Large Language Models (LLMs) in annotating Dutch assessment questions to support curriculum alignment. To evaluate this, it was used a dataset of 4,000 Dutch questions from university-level statistics courses. The study utilized both generative (GPT-4.1 nano) and non-generative (mBERT) models, comparing their performance in classifying questions based on subject-specific concepts of statistics from a predefined hierarchical taxonomy. The research also highlighted some of the key factors that influence their effectiveness in this specialized task.

Overall, LLMs demonstrated strong performance and promising potential in annotating Dutch assessment questions. These findings directly address the study's research questions (RQ1), confirming that LLMs can support educators by reducing the time and effort required to craft and evaluate questions, allowing them to focus more on promoting cognitive engagement. By ensuring that assessments are better aligned with learning objectives, educators can enhance both the effectiveness of their teaching and the overall learning experience for students.

5.1 Prompt Engineering Impact

Addressing the first subquestion (SQ1), this study demonstrates that prompt format and the inclusion of contextual information affect model outputs. Specifically, this study found that carefully elaborating the prompt is crucial for optimizing GPT's performance in educational annotation tasks, and that effective use of LLMs requires adapting the prompt design to suit the task's nature. For instance, using a predefined set of labels, prompts that included step-by-step instructions (**Chain-of-Thought**) or simulated a collaborative discussion between three domain experts (**Experts**) performed best for annotation. In contrast, **Few-shot** learning, which included three examples of annotated questions, achieved better results in generative tasks where no predefined list was provided, guiding the model toward more accurate and contextually aligned topic generation.

Although this research was conducted in a multilingual context, where performance typically declines due to LLMs being primarily trained on English data, explicitly specifying Dutch in the prompt resulted in noticeable improvements in both type of tasks. This suggests that prompting the model to act as a Dutch expert enables it to generate more accurate and contextually appropriate annotations, reflecting a deeper, language-aware understanding of the questions. As previously demonstrated by H. Huang et al. (2023) and Xu et al. (2025), this reinforces the growing consensus that multilingual prompt engineering is a valuable strategy for improving LLM performance in low-resource and educational settings.

5.1.1 Classification with a predefined list of labels

To evaluate and compare with the true labels, the model was required to select a subject from a predefined list of statistical concepts. The results showed that including one or more examples in the prompt (**One** and **Few-shot**) did not consistently improve model performance in this context, especially when measured by macro-averaged F1 scores. In one-shot learning, performance improved for the category used in the example, but the model failed to generalize to other categories, likely due to differences in topic granularity. As shown in the exploratory analysis (Table 4.2), some categories such as Descriptive Statistics have few subtopics and are relatively simple, while others, such as Inferential Statistics and Factor Analysis, are more complex and diverse.

Interestingly, increasing the number of examples in the prompt led to a decline in both accuracy and macro F1 score. When manually evaluating some examples where GPT provides a justification of its label choice for each question, the **Few-shot** prompt showed to be overly influenced by the examples provided during training, highlighting the method's sensitivity to example selection. Overall, this pattern supports findings from previous studies (Faraby et al., 2024; Li et al., 2024), indicating that in some cases **Few-shot** learning may introduce additional complexity, potentially due to increased prompt length or conflicting contextual examples.

In contrast, prompts designed to encourage step-by-step reasoning, such as the **CoT** and **Experts** approaches, led to relevant performance gains, with up to a 10% increase in accuracy and a 13.9% improvement in macro-averaged F1 score when a predefined list of labels were provided. The manual evaluation of some outputs highlighted that step-by-step prompting enables the model to reason more thoroughly and avoid superficial keyword associations, an issue often observed with the **Baseline** and **Few-shot** approaches, resulting in more accurate topic identification.

However, while these reasoning-based prompts led to improved overall performance, the results also indicate that GPT's classification varies across the categories. When analyzing the results of **CoT** and **Experts** approach by main-level (shown in Table 4.3), there is a clear difference in performance for some subjects. While Inferential Statistics achieved an 83.5% F1 score in **CoT** approach, Distributions category achieved only 25.2%, demonstrating to be a challenging category to categorize. This difficulty could stem from several factors, such as the nuanced differences within this category or perhaps a reliance on graphical interpretation in many distribution-related questions that were not included in the prompt. Therefore, this highlights that although GPT can achieve strong results, its performance can vary across different types of questions, underscoring the need for careful evaluation when applying it to diverse educational contexts.

5.1.2 Generative Prediction Classification

Building on these insights into prompt engineering for fixed-label classification, this study also explored a distinct prompting strategy: instructing the model to generate topic labels without relying on a predefined list. While the earlier classification task focused on broader categorization using top-level labels from a fixed taxonomy, this generative setup enabled a more granular analysis of the model's ability to articulate specific statistical concepts. Interestingly, the optimal prompting strategies for this generative task diverged from those found effective in the fixed-label classification, as will be discussed in this section.

Unlike in the classification task, where reasoning-based strategies such as **CoT** and **Experts** outperformed the simpler baseline prompt, these approaches did not yield strong results in the generative setting. Given that the task now required not only interpreting and annotating the question but also generating a label, these strategies may have introduced unnecessary complexity. In contrast, the **Few-shot** prompting strategy proved to be more effective in guiding the model toward accurate and semantically aligned topic generation. This was reflected in improved F1 scores across several categories, particularly the ones included in the example set, underscoring the model's capacity to generalize from well-annotated instances.

This performance compares favorably with prior research. For example, Moore et al. (2024) explored labeling educational questions without predefined labels. Their best-performing LLM strategy achieved human-evaluated concept matches of 56% for Chemistry and 35% for E-Learning multiple-choice questions. While their task involved different domains, this comparison highlights the relative effectiveness of a well-constructed few-shot strategy in guiding LLMs toward conceptually accurate outputs in open-ended annotation tasks.

Analyses of the other different approaches tested (**Textbook** and **Experts** described in Section 3.2.1) show that both performed worse than the Baseline. The Textbook prompt often produced overly broad labels resembling section headers rather than specific concepts, suggesting that prompts designed around locating content within large knowledge sources may lack the specificity needed for fine-grained annotation tasks. On the other hand, the **Experts** prompt frequently produced responses framed in terms of reasoning or procedural actions. Although often semantically related, this highlights a challenge in balancing the richness of generated explanations with the precision required for topic assignment. Outputs that are too abstract or generalized may be valuable for understanding but problematic for tasks demanding proper annotation of a question.

These results highlight the importance of tailoring prompting strategies to the nature of the task. While step-by-step approaches enhance performance in classification tasks with fixed label sets by encouraging deeper reasoning, they may reduce precision in open-ended generation tasks where concise concept-specific output is essential. Conversely, giving some examples in the instruction (**Few-shot**), demonstrated strong potential to guide LLMs toward more accurate and contextually aligned topic generation in generative labeling tasks.

5.2 Comparison Between GPT and mBERT

Having examined the nuances of prompting strategies for generative models in both fixed-label and open-ended annotation tasks, a crucial next step was to contextualize their performance against non-generative approaches, thereby addressing the second subquestion (SQ2). Therefore, a comparison with the best-performing generative prompt to a supervised fine-tuned mBERT model was conducted, revealing significant differences in their capabilities and limitations.

When comparing the models, mBERT demonstrated substantially superior performance, achieving a 27.8% higher accuracy (91.7% vs. 71.7%) and a 34.5% increase in macro-averaged F1 score (83.7% vs. 62.2%). This advantage can be attributed to the greater control and adaptability afforded by the supervised learning framework. While GPT-based models achieved comparable results, their internal decision-making processes are largely opaque, limiting interpretability and constraining error analysis or targeted improvements. In contrast, mBERT allows hyperparameter tuning and the application of strategies to mitigate challenges such as class imbalance, which was relevant in this study. Furthermore, a sufficiently large dataset enabled a robust train–validation–test split, supporting reliable model optimization.

Extending the evaluation to the second level of the hierarchical taxonomy revealed a limitation in GPT's performance and reinforced the advantage of the fine-tuned mBERT model. This drop is likely due to the increased semantic proximity among second-level labels, which fall under the same broad top-level categories and therefore often share overlapping terminology and conceptual content, making them more difficult for generative models to differentiate without explicit supervision. Therefore, as highlighted by Mizrahi et al. (2024) and Xu et al. (2025), the robustness of LLMs to prompt variations remains a challenge and should be given careful attention for automation using generative models.

These findings carry important implications for educators and developers selecting AI models for automating curriculum alignment. On one hand, GPT-based approaches are particularly suitable in contexts without labeled training data due to their strong zero-shot capabilities. Additionally, they require less technical expertise and computational resources, as they are accessible through APIs without the need for local infrastructure or training. Their flexibility also makes them well-suited for tasks without fixed label sets, supporting exploratory applications. However, GPT models incur recurring operational costs tied to commercial API usage, making them less feasible for educational institutions or researchers with limited budgets. Moreover, their performance is highly dependent on the quality and structure of the prompts provided, often requiring extensive prompt engineering and iterative refinement. Furthermore, their robustness across different datasets or domains remains uncertain, especially in applications that demand high reliability or domain-specific consistency.

On the other hand, fine-tuned BERT models are preferable when extensive labeled data is available and the task involves classification into fixed categories. These models are open-source, allowing control over data handling, model tuning, and optimization for issues such as class imbalance. In addition, BERT models do not require extra costs during deployment and offer greater transparency during training and evaluation, enabling more explainability of its results. However, the effectiveness of supervised approaches is heavily dependent on the quality and quantity of the training data. These methods typically require access to powerful GPUs and involve more effort in terms of text preprocessing, experimentation with hyperparameters, and implementation. In educational contexts, where annotated datasets are often limited or expensive to produce, these requirements may pose significant barriers.

In summary, while GPT-based models offer a lightweight and accessible solution for many applications, especially in low-resource settings, their performance is constrained by their sensitivity to prompt formulation and the lack of interpretability. On the other hand, BERT-based models offer greater control and performance in high-data settings but require more substantial infrastructure and expertise. The decision between these approaches should be guided by the specific goals of the task, the availability of resources, and the nature of the data involved.

5.3 Limitations and future work

While this study demonstrated the potential of Large Language Models (LLMs) for automating curriculum annotation, several limitations emerged that highlight important directions for future research.

Regarding the GPT application, there were some limitations due to cost constraints. First, variability of the model's output was approximated using the baseline's observed standard deviation rather than by generating multiple responses per prompt. While this proxy were used to assess meaningful improvements beyond model variability, future work should more rigorously explore temperature settings and replicate prompts multiple times to better isolate the effects of prompt design from model stochasticity. In addition, the choice of the model (GPT 4.1-nano) was also made considering lower costs. Although this model demonstrates strong performance, OpenAI offers larger and more powerful models that could also be applied in this context and explored in future work.

When analyzing the results of different prompts, those that included examples performed worse when a predefined list of topics was provided. Although this result aligns with findings from other studies, the effectiveness of this approach may be limited by the selection of examples. Prior research (Wei et al., 2023) emphasizes that few-shot performance depends heavily on the quality of examples and the complexity of the task. Although the examples used here were carefully chosen, future studies could adopt more systematic or data-driven example selection to enhance generalization and robustness.

In terms of the dataset used in this research, the hierarchical taxonomy created by Namesnik-Silvester et al. (2025) was only partially explored. While generative prediction classification helped address deeper taxonomy levels, the classification task with a predefined list of labels, as well as the mBERT model, was not tested across multiple hierarchy levels due to data limitations. Future work could, for instance, incorporate strategies like active learning (van Grinsven et al., 2023) that help optimize label selection in domains where data is costly to obtain. Additionally, exploring alternative datasets without strict hierarchies or designing prompts and training data tailored to fine-grained hierarchical classification could capture more nuanced distinctions between concepts, enabling richer conceptual detail analysis.

Furthermore, this study focused exclusively on annotation of question within the domain of Statistics. Future research could investigate the applicability of LLMs to other subject areas, such as biology, physics, or language education, to evaluate the generalizability of the strategies used in this study. In addition, it would also be valuable for curriculum alignment evaluations to expand beyond subject concept annotation and explore other relevant categories, such as learning levels or cognitive processes, to provide a more comprehensive understanding of assessment content and its alignment with educational objectives.

Beyond dataset and domain scope considerations, further refinement of prompting strategies remains a key avenue for improving performance. For instance, incorporating the full solution to each question or including metadata such as the cognitive level being assessed could help models more accurately infer the underlying concepts. Additionally, regarding second-level classification, this study was constrained by time and did not explore prompts specifically tailored to the narrower distinctions within subcategories. Future research could investigate more context-aware prompting strategies or hierarchical prompt designs to improve performance on fine-grained educational classification tasks.

To conclude, while cosine similarity and GPT-based evaluation provided a practical alternative for assessing generative classification performance, human evaluation would offer a more reliable assessment. However, incorporating expert judgment was beyond the scope of this study due to the substantial resources and coordination it would require. The observed discrepancies between the two evaluation methods highlight a key challenge in assessing open-ended generative outputs: embedding-based metrics like cosine similarity may undervalue predictions that are semantically correct but phrased differently, particularly when reference labels are hierarchical or highly detailed. This underscores the need for evaluation approaches that better capture meaning beyond surface-level similarity. As such, further validation of automated methods is necessary, especially given their sensitivity to prompt design. Future work could, for instance, explore ways to calibrate GPT-based evaluations against expert ratings, improving the reliability and interpretability of automated assessment in educational settings.

5.4 Concluding Remarks

This research investigated how well large language models annotate Dutch assessment questions with curriculum subject-specific concepts, which is a crucial aspect to support curriculum alignment applications and, therefore, enhance diverse educational learning process. By exploring different prompt approaches and understanding in which ways do GPT and mBERT are different in annotating questions, this study shows the effectiveness in this type of task without requiring extensive expert annotations and enabling automatic applications. Therefore, this study contributes to the integration of LLMs into education, offering an effective solution for automating annotation questions, covering which approach is more applicable in several scenarios in education applications.

Appendix

This appendix provides supplementary materials that support the main findings and discussions presented in the thesis. It includes prompt templates and additional outputs that are referenced throughout the study.

The code used in this research is available at the following GitHub repository: https://github.com/isahlm/llm-dutch-question-annotation

Table A.1: All prompts format tested in In Context Learning trials.

Prompt	**Content**
Baseline	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. User: Analyze the following question to determine the most appropriate topic from this list of statistical concepts: ['assumptions', 'factor analysis', 'descriptive statistics', 'inferential statistics', 'distributions', 'reliability', 'type of variable', 'probability']. Your task is to examine the question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Question: {input} Please provide both the topic and rationale as follows: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic>
Baseline + Cross Lingual	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. You should act as an expert in understanding and interpreting questions written in Dutch. User: Analyze the following Dutch question to determine the most appropriate topic from this list of statistical concepts: ['assumptions', 'factor analysis', 'descriptive statistics', 'inferential statistics', 'distributions', 'reliability', 'type of variable', 'probability']. Your task is to examine the Dutch question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Question: {input} Important: Although the question is in Dutch, your output must be in English. Please provide both the topic and rationale as follows: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic>

Prompt	**Content**
Baseline + Cross Lingual + OneShot	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. You should act as an expert in understanding and interpreting questions written in Dutch. User: Analyze the following Dutch question to determine the most appropriate topic from this list of statistical concepts: [’assumptions’, ’factor analysis’, ’descriptive statistics’, ’inferential statistics’, ’distributions’, ’reliability’, ’type of variable’, ’probability’]. Your task is to examine the Dutch question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Example: Question: Hieronder is van 5 individuen aangegeven hoe zij t.o.v. elkaar gerangschikt zijn op autoritair gedrag en sociale status: Autoritair gedrag: 3, 5, 1, 2, 4 Sociale status: 2, 3, 1, 4, 5. De waarde van de Spearman correlatiecoëfficiënt is hier Answerlist ———- * -0.5 * 0.4 * 0.5 * 0.513 Final Topic: descriptive statistics Rationale: This question involves calculating the Spearman correlation coefficient, which is a statistical measure used to summarize the strength and direction of a relationship between two ranked variables. It describes patterns in the given data without making predictions or generalizations beyond it. Now, analyze the Dutch question below. Question: {input} Important: Although the question is in Dutch, your output must be in English. Please provide both the topic and rationale in as follows: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic>

Prompt	**Content**
Baseline + Cross Lingual + FewShot	Here same System and User used in One-Shot Prompt, but additional examples as follows: [...] Example 1: Question: Hieronder is van 5 individuen aangegeven hoe zij t.o.v. elkaar gerangschikt zijn op autoritair gedrag en sociale status: Autoritair gedrag: 3, 5, 1, 2, 4 Sociale status: 2, 3, 1, 4, 5. De waarde van de Spearman correlatiecoëfficiënt is hier Answerlist ———- * -0.5 * 0.4 * 0.5 * 0.513 Final Topic: descriptive statistics Rationale: This question involves calculating the Spearman correlation coefficient, which is a statistical measure used to summarize the strength and direction of a relationship between two ranked variables. It describes patterns in the given data without making predictions or generalizations beyond it. Example 2: Question: Het gemiddelde in een steekproef van 500 personen is 5 en de variantie is 1. Bereken het 95% betrouwbaarheidsinterval voor het populatiegemiddelde. Answerlist ———- * 4.93; 5.07 * 4.91; 5.09 * 3.36; 6.65 * 3.04; 6.96 Final Topic: inferential statistics Rationale: This question asks to calculate a 95% confidence interval for the population mean based on a sample. Inferential statistics involves using sample data to make estimates or draw conclusions about a larger population. Example 3: Question: Onderstaand Venn-diagram geeft weer welk deel van de variantie in y verklaard wordt door de twee predictoren x1 en x2. Welke conclusie over het Venn-diagram is juist? Answerlist ———- - Er is nauwelijks sprake van multicollineariteit. - Er is sprake van een lage coëfficiënt van multipele determinatie R2. - Er is sprake van een lage multipele correlatie R. - Er is sprake van aanzienlijke interactie. Final Topic: assumptions Rationale: This question focuses on interpreting the relationships between predictors and their explained variance in a regression context, specifically addressing issues like multicollinearity and interaction. These relate directly to the assumptions underlying multiple regression models, which must hold true for valid results. Now, analyze the Dutch question below. [...]

Prompt	Content
Baseline + Cross Lingual + Chain-of-Thoughts	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. You should act as an expert in understanding and interpreting questions written in Dutch. User: Analyze the following Dutch question to determine the most appropriate topic from this list of statistical concepts: [’assumptions’, ’factor analysis’, ’descriptive statistics’, ’variable type’, ’inferential statistics’, ’distributions’, ’reliability’, ’measurement level’, ’probability’]. Your task is to examine the Dutch question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Internally, think step-by-step as follows: Step 1: Read and understand the Dutch question. Summarize its main statistical focus in English. Step 2: Identify key statistical terms or concepts implicit or explicit in the question. Step 3: Match these terms to the most appropriate topic from the given list, highlighting why the selected topic fits best and why others do not. Step 4: Provide a detailed rationale explaining why you selected this topic. Important: Although the question is in Dutch, your output must be in English. Output ONLY the following: Final Topic: <exact topic from the list> Rationale: <detailed explanation for the selected topic> Question: {input}

Table A.2: All prompts format tested when including Experts and Textbook variation Prompts

Prompt	Content
Baseline	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. User: Analyze the following question to determine the most appropriate topic from this list of statistical concepts: [’assumptions’, ’factor analysis’, ’descriptive statistics’, ’inferential statistics’, ’distributions’, ’reliability’, ’type of variable’, ’probability’]. Your task is to examine the question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Question: {input} Please provide both the topic and rationale as follows: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic>

Prompt	**Content**
Experts	System: You are simulating a panel of three expert statistics instructors who are collaboratively analyzing student assessment questions. Each expert is brilliant, logical, detail-oriented, and highly critical. Their task is to discuss and determine what statistical knowledge components and cognitive skills the question assesses. The discussion should reflect thoughtful academic reasoning and rigorous pedagogical analysis. User: The following question is part of a low-stakes assessment in a university-level statistics course. Simulate a collaborative discussion among three expert instructors. Each expert should interpret the question, then articulate their reasoning in detail and in real time, referencing one another, asking questions, making corrections, and building consensus. The conversation should reflect a deep analysis of the question, focusing on what knowledge and skills it is testing. The experts must categorize the question into one topic from the following rubric: ['assumptions', 'factor analysis', 'descriptive statistics', 'inferential statistics', 'distributions', 'reliability', 'type of variable', 'probability']. You must not output the discussion, but ONLY the following: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic> Question: {input}
Textbook	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. User: Below there is a question written intended for a university-level audience with existing prior knowledge on the subject of Statistics. The question is used as a low-stakes assessment as part of an Statistics course that covers similar content. If this question was presented in a textbook for an Statistics course, what single domain-specific topic from the following list would the page cover? This should be based on the knowledge components and skills the question assesses. The possible topics to choose from are: ['assumptions', 'factor analysis', 'descriptive statistics', 'inferential statistics', 'distributions', 'reliability', 'type of variable', 'probability']. Please provide both the topic and rationale as follows: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic> Question: {input}

Prompt	**Content**
Experts + Cross Lingual	System: You are simulating a panel of three expert statistics instructors who are collaboratively analyzing student assessment questions and are expert in understanding and interpreting questions written in Dutch. Each expert is brilliant, logical, detail-oriented, and highly critical. Their task is to discuss and determine what statistical knowledge components and cognitive skills the question assesses. The discussion should reflect thoughtful academic reasoning and rigorous pedagogical analysis. User: The following question is written in Dutch and is part of a low-stakes assessment in a university-level statistics course. Simulate a collaborative discussion among three expert instructors. Each expert should interpret the Dutch question, then articulate their reasoning in detail and in real time, referencing one another, asking questions, making corrections, and building consensus. The conversation should reflect a deep analysis of the question, focusing on what knowledge and skills it is testing. The experts must categorize the question into one topic from the following rubric: ['assumptions', 'factor analysis', 'descriptive statistics', 'inferential statistics', 'distributions', 'reliability', 'type of variable', 'probability']. Important: Although the question is in Dutch, your output must be in English. You must not output the discussion, but ONLY the following: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic> Question: {input}

Prompt	**Content**
Experts + Cross Lingual + One Shot	System: You are simulating a panel of three expert statistics instructors who are collaboratively analyzing student assessment questions and are expert in understanding and interpreting questions written in Dutch. Each expert is brilliant, logical, detail-oriented, and highly critical. Their task is to discuss and determine what statistical knowledge components and cognitive skills the question assesses. The discussion should reflect thoughtful academic reasoning and rigorous pedagogical analysis. User: The following question is written in Dutch and is part of a low-stakes assessment in a university-level statistics course. Simulate a collaborative discussion among three expert instructors. Each expert should interpret the Dutch question, then articulate their reasoning in detail and in real time, referencing one another, asking questions, making corrections, and building consensus. The conversation should reflect a deep analysis of the question, focusing on what knowledge and skills it is testing. The experts must categorize the question into one topic from the following rubric: ['assumptions', 'factor analysis', 'descriptive statistics', 'inferential statistics', 'distributions', 'reliability', 'type of variable', 'probability']. Important: Although the question is in Dutch, your output must be in English. You must not output the discussion, but ONLY the following: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic> Example: Question: Hieronder is van 5 individuen aangegeven hoe zij t.o.v. elkaar gerangschikt zijn op autoritair gedrag en sociale status: Autoritair gedrag: 3, 5, 1, 2, 4 Sociale status: 2, 3, 1, 4, 5. De waarde van de Spearman correlatiecoëfficiënt is hier Answerlist ———- * -0.5 * 0.4 * 0.5 * 0.513 Final Topic: descriptive statistics Rationale: This question involves calculating the Spearman correlation coefficient, which is a statistical measure used to summarize the strength and direction of a relationship between two ranked variables. It describes patterns in the given data without making predictions or generalizations beyond it. Now, analyze the Dutch question below: Question: input

Prompt	**Content**
Experts + Cross Lingual + Few Shot	Here same System and User used in One-Shot Prompt, but additional examples as follows: [...] Example 1: Question: Hieronder is van 5 individuen aangegeven hoe zij t.o.v. elkaar gerangschikt zijn op autoritair gedrag en sociale status: Autoritair gedrag: 3, 5, 1, 2, 4 Sociale status: 2, 3, 1, 4, 5. De waarde van de Spearman correlatiecoëfficiënt is hier Answerlist ———- * -0.5 * 0.4 * 0.5 * 0.513 Final Topic: descriptive statistics Rationale: This question involves calculating the Spearman correlation coefficient, which is a statistical measure used to summarize the strength and direction of a relationship between two ranked variables. It describes patterns in the given data without making predictions or generalizations beyond it. Example 2: Question: Het gemiddelde in een steekproef van 500 personen is 5 en de variantie is 1. Bereken het 95% betrouwbaarheidsinterval voor het populatiegemiddelde. Answerlist ———- * 4.93; 5.07 * 4.91; 5.09 * 3.36; 6.65 * 3.04; 6.96 Final Topic: inferential statistics Rationale: This question asks to calculate a 95% confidence interval for the population mean based on a sample. Inferential statistics involves using sample data to make estimates or draw conclusions about a larger population. Example 3: Question: Onderstaand Venn-diagram geeft weer welk deel van de variantie in y verklaard wordt door de twee predictoren x1 en x2. Welke conclusie over het Venn-diagram is juist? Answerlist ———- * Er is nauwelijks sprake van multicollineariteit. * Er is sprake van een lage coëfficiënt van multipele determinatie R^2. $* Erissprakevaneenlagemultipelecorrelatie R. * Erissprakevanaanzienlijkeinteractie.$ Final Topic: assumptions Rationale: This question focuses on interpreting the relationships between predictors and their explained variance in a regression context, specifically addressing issues like multicollinearity and interaction. These relate directly to the assumptions underlying multiple regression models, which must hold true for valid results. Now, analyze the Dutch question below: [...]

Table A.3: All prompts format tested in Generative Prediction Classification task (no predefined list of labels provided)

Prompt	**Content**
Baseline	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. User: Analyze the following question to determine the most appropriate topic of statistical concepts. Your task is to examine the question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Please provide both the topic and rationale as follows: Final Topic: <topic> Rationale: <detailed explanation for the selected topic> Question: {input}
Experts	System: You are simulating a panel of three expert statistics instructors who are collaboratively analyzing student assessment questions. Each expert is brilliant, logical, detail-oriented, and highly critical. Their task is to discuss and determine what statistical knowledge components and cognitive skills the question assesses. The discussion should reflect thoughtful academic reasoning and rigorous pedagogical analysis. User: The following question is part of a low-stakes assessment in a university-level statistics course. Simulate a collaborative discussion among three expert instructors. Each expert should interpret the question, then articulate their reasoning in detail and in real time, referencing one another, asking questions, making corrections, and building consensus. The conversation should reflect a deep analysis of the question, focusing on what knowledge and skills it is testing. You must not output the discussion, but ONLY the following: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic> Question: {input}
Textbook	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. User: Below there is a question written intended for a university-level audience with existing prior knowledge on the subject of Statistics. The question is used as a low-stakes assessment as part of an Statistics course that covers similar content. If this question was presented in a textbook for an Statistics course, what single domain-specific topic from the following list would the page cover? This should be based on the knowledge components and skills the question assesses. Please provide both the topic and rationale as follows: Final Topic: <exact topic from the list above> Rationale: <detailed explanation for the selected topic> Question: {input}

Prompt	**Content**
Baseline + Cross Lingual	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. You should act as an expert in understanding and interpreting questions written in Dutch. User: Analyze the following Dutch question to determine the most appropriate topic of statistical concepts. Your task is to examine the Dutch question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Important: Although the question is in Dutch, your output must be in English. Please provide both the topic and rationale as follows: Final Topic: <topic> Rationale: <detailed explanation for the selected topic> Question: {input}
Baseline + Cross Lingual + One Shot	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. You should act as an expert in understanding and interpreting questions written in Dutch. User: Analyze the following Dutch question to determine the most appropriate topic of statistical concepts. Your task is to examine the Dutch question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Important: Although the question is in Dutch, your output must be in English. Please provide both the topic and rationale as follows: Final Topic: <topic> Rationale: <detailed explanation for the selected topic> Example: Question: Hieronder is van 5 individuen aangegeven hoe zij t.o.v. elkaar gerangschikt zijn op autoritair gedrag en sociale status: Autoritair gedrag: 3, 5, 1, 2, 4 Sociale status: 2, 3, 1, 4, 5. De waarde van de Spearman correlatiecoëfficiënt is hier Answerlist ———- * -0.5 * 0.4 * 0.5 * 0.513 Final Topic: descriptive statistics Rationale: This question involves calculating the Spearman correlation coefficient, which is a statistical measure used to summarize the strength and direction of a relationship between two ranked variables. It describes patterns in the given data without making predictions or generalizations beyond it. Now, analyze the Dutch question below: Question: {input}

Prompt	**Content**
Baseline + Cross Lingual + Few Shot	Here same System and User used in One-Shot Prompt, but additional examples as follows: [...] Example 1: Question: Hieronder is van 5 individuen aangegeven hoe zij t.o.v. elkaar gerangschikt zijn op autoritair gedrag en sociale status: Autoritair gedrag: 3, 5, 1, 2, 4 Sociale status: 2, 3, 1, 4, 5. De waarde van de Spearman correlatiecoëfficiënt is hier Answerlist ———- * -0.5 * 0.4 * 0.5 * 0.513 Final Topic: descriptive statistics Rationale: This question involves calculating the Spearman correlation coefficient, which is a statistical measure used to summarize the strength and direction of a relationship between two ranked variables. It describes patterns in the given data without making predictions or generalizations beyond it. Example 2: Question: Het gemiddelde in een steekproef van 500 personen is 5 en de variantie is 1. Bereken het 95% betrouwbaarheidsinterval voor het populatiegemiddelde. Answerlist ———- * 4.93; 5.07 * 4.91; 5.09 * 3.36; 6.65 * 3.04; 6.96 Final Topic: inferential statistics Rationale: This question asks to calculate a 95% confidence interval for the population mean based on a sample. Inferential statistics involves using sample data to make estimates or draw conclusions about a larger population. Example 3: Question: Onderstaand Venn-diagram geeft weer welk deel van de variantie in y verklaard wordt door de twee predictoren x1 en x2. Welke conclusie over het Venn-diagram is juist? Answerlist ———- * Er is nauwelijks sprake van multicollineariteit. * Er is sprake van een lage coëfficiënt van multipele determinatie R^2. $* ErissprakevaneenlagemultipelecorrelatieR. *$ $Erissprakevanaanzienlijkeinteractie.$ Final Topic: assumptions Rationale: This question focuses on interpreting the relationships between predictors and their explained variance in a regression context, specifically addressing issues like multicollinearity and interaction. These relate directly to the assumptions underlying multiple regression models, which must hold true for valid results. Now, analyze the Dutch question below: [...]

Prompt	**Content**
Baseline + Cross Lingual + Chain-of-Thoughts	System: You are an expert in educational content evaluation, specializing in university-level Statistics. You classify questions according to core statistical concepts. You should act as an expert in understanding and interpreting questions written in Dutch. User: Analyze the following Dutch question to determine the most appropriate topic of statistical concepts. Your task is to examine the Dutch question text below and select the most fitting topic based on the knowledge components and skills the question assesses. Internally, think step-by-step as follows: Step 1: Read and understand the Dutch question. Summarize its main statistical focus in English. Step 2: Identify key statistical terms or concepts implicit or explicit in the question. Step 3: Match these terms to the most appropriate topic from the given list, highlighting why the selected topic fits best and why others do not. Step 4: Provide a detailed rationale explaining why you selected this topic. Important: Although the question is in Dutch, your output must be in English. Please provide both the topic and rationale as follows: Final Topic: <topic> Rationale: <detailed explanation for the selected topic> Question: {input}

Tables A.4, A.5 and A.6 helped evaluate the performance of the prompts per main-level category.

Table A.4: F1 Score (%) by Class and Number of Samples across Prompting Strategies for In-Context Learning Results

Class	Baseline	+Cross Lingual	+Cross Lingual + One Shot	+Cross Lingual + Few-Shot	+Cross Lingual + CoT	Number of Samples
assumptions	50.8%	60.7%	33.3%	19.6%	60.0%	19
descriptive statistics	58.1%	56.8%	64.6%	65.7%	57.5%	119
distributions	22.0%	22.3%	24.3%	20.9%	25.2%	42
factor analysis	42.9%	60.0%	66.7%	54.6%	66.7%	8
inferential statistics	78.6%	79.8%	80.8%	60.8%	83.5%	483
probability	68.0%	62.6%	71.5%	66.7%	71.3%	95
reliability	50.0%	63.2%	58.8%	66.7%	66.7%	10
type of variable	66.7%	66.7%	63.2%	77.8%	66.7%	8
macro avg	54.6%	59.0%	57.9%	54.1%	62.2%	784

Table A.5: F1 Score (%) by Class and Number of Samples across Prompting Strategies including Textbook and Experts approaches

Class	Baseline	Experts	Textbook	Experts + Cross Lingual	Experts + Cross Lingual + OneShot	Experts + Cross Lingual + Fewshot	Number of Samples
assumptions	50.8%	64.0%	50.8%	65.4%	42.1%	21.1%	19
descriptive statistics	58.1%	56.0%	60.8%	57.1%	59.0%	65.5%	119
distributions	22.0%	23.0%	25.0%	24.6%	25.0%	21.0%	42
factor analysis	42.9%	61.0%	60.0%	57.1%	70.0%	52.0%	8
inferential statistics	78.6%	82.0%	79.0%	83.2%	82.0%	67.0%	483
probability	67.9%	71.0%	72.0%	71.4%	69.0%	72.0%	95
reliability	50.0%	57.0%	63.0%	60.0%	63.0%	53.0%	10
type of variable	66.7%	63.0%	64.0%	66.7%	53.0%	70.0%	8
macro avg	54.6%	60.0%	59.0%	60.7%	58.0%	53.0%	784

Table A.6: Accuracy (%) by Class and Number of Samples across Generative Classification Prompting Strategies

Class	Baseline	Experts	Textbook	Baseline + CL	Baseline + CL + OneShot	Baseline + CL + FewShot	Baseline + CL + CoT	Number of Samples
assumptions	42.1%	57.9%	42.1%	57.9%	15.8%	94.7%	42.1%	19
descriptive statistics	55.5%	54.6%	53.8%	53.8%	54.6%	60.5%	51.3%	119
distributions	40.5%	52.4%	47.6%	42.9%	35.7%	16.7%	42.9%	42
factor analysis	87.5%	75.0%	75.0%	87.5%	87.5%	62.5%	87.5%	8
inferential statistics	81.2%	71.8%	76.2%	80.8%	81.6%	83.9%	77.2%	483
probability	59.0%	52.6%	55.8%	55.8%	63.2%	65.3%	55.8%	95
reliability	50.0%	50.0%	50.0%	60.0%	40.0%	50.0%	70.0%	10
type of variable	50.0%	62.5%	25.0%	87.5%	75.0%	87.5%	62.5%	8

The Tables A.7, A.8,A.9 and A.10 present the results of f1 score per second-level categories.

Table A.7: F1-score comparison for subcategories of Inferential Statistics

Subcategory	mBERT	GPT	Support
Bayesian statistics	85.7%	66.7%	4
Confidence intervals	72.7%	35.1%	13
Effect size	61.5%	57.1%	11
Multilevel analysis	40.0%	44.4%	2
NHST	76.1%	58.9%	93
Non-parametric techniques	94.7%	8.7%	20
Parametric techniques	77.8%	35.9%	169
Regression	88.5%	87.1%	129
Sampling distributions	62.1%	33.3%	12

Table A.8: F1-score comparison for subcategories of Descriptive Statistics

Subcategory	mBERT	GPT	Support
Data representation	81.8%	53.3%	19
Score interpretation	85.7%	34.3%	8
Summary statistics	93.8%	73.9%	82

Table A.9: F1-score comparison for subcategories of Probability

Subcategory	mBERT	GPT	Support
Conditional probability	47.1%	23.1%	4
Events	50.0%	27.0%	5
General rules	83.1%	11.1%	32
Random variables	80.8%	48.7%	27
Sample space	93.3%	57.1%	8

Table A.10: F1-score comparison for subcategories of Distributions

Subcategory	mBERT	GPT	Support
Continuous	94.1%	50.0%	33
Discrete	83.3%	40.0%	7
Limit theorems	0.0%	40.0%	2

Bibliography

Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory into Practice*, *41*(4), 255–260. https://doi.org/10.1207/s15430421tip4104_9

Blom, B., & Pereira, J. L. (2023). Domain adaptation in transformer models: Question answering of dutch government policies. *International Conference on Intelligent Data Engineering and Automated Learning*, 196–208. https://doi.org/10.1007/978-3-031-48232-8_19

Chae, Y., & Davidson, T. (2024). Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods Research*. https://doi.org/10.1177/00491241251325243

Faraby, S. A., Romadhony, A., & Adiwijaya. (2024). Analysis of llms for educational question classification and generation. *Computers and Education: Artificial Intelligence*, *7*, 100298. https://doi.org/10.1016/j.caeai.2024.100298

Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., & Wei, F. (2023). Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *Findings of the Association for Computational Linguistics: EMNLP 2023*. https://doi.org/10.18653/v1/2023.findings-emnlp.826

Huang, T., Yang, H., Hu, S., Geng, J., Liu, S., Zhang, H., & Yang, Z. (2023). Pqsct: Pseudo-siamese bert for concept tagging with both questions and solutions. *IEEE Transactions on Learning Technologies*, *16*(5), 831–846. https://doi.org/10.1109/TLT.2023.3275707

Isbister, T., Carlsson, F., & Sahlgren, M. (2021). Should we stop training more monolingual models, and simply use machine translation instead? *CoRR*. https://doi.org/10.48550/arXiv.2104.10441

Kwak, Y., & Pardos, Z. A. (2024). Bridging large language model disparities: Skill tagging of multilingual educational content. *British Journal of Educational Technology*, *55*(5), 2039–2057. https://doi.org/10.1111/bjet.13465

Li, H., Xu, T., Tang, J., & Wen, Q. (2024). Automate knowledge concept tagging on math questions with llms. *ArXiv*. https://doi.org/10.48550/arXiv.2403.17281

Liu, C., Zhang, W., Zhao, Y., Luu, A. T., & Bing, L. (2024). Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv*. https://doi.org/10.48550/arXiv.2403.10258

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, *12*, 933–949. https://doi.org/10.48550/arXiv.2401.00595

Moore, S., Schmucker, R., Mitchell, T., & Stamper, J. (2024). Automated generation and tagging of knowledge components from multiple-choice questions. *Proceedings of the Eleventh ACM Conference on Learning at Scale*, 122–133. https://doi.org/10.48550/arXiv.2405.20526

Mustafidah, H., Suwarsito, S., & Pinandita, T. (2022). Natural language processing for mapping exam questions to the cognitive process dimension. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(13), 4–16. https://doi.org/10.3991/ijet.v17i13.29095

Namesnik-Silvester, K., Polak, M., Smits, N., Swinkels, J., Arends, L., Pavlopoulos, D., Psychogyiopoulos, A., Lindemann, O., Klinkenberg, S., & de Moor, M. H. (2025). Sharestats: An open statistics item bank developed by a community of instructors in higher education. *Teaching Statistics*, *47*(2), 129–138. https://doi.org/10.1111/test.12399

Nappi, J. S. (2017). The importance of questioning in developing critical thinking skills. *Delta Kappa Gamma Bulletin*, *84*(1), 30.

Osman, A., & Yahya, A. A. (2016). Classifications of exam questions using natural language syntactic features: A case study based on bloom's taxonomy. *Proc. 3rd Int. Arab Conf. Qual. Assurance Higher Educ*, 1–8.

Qin, L., Chen, Q., Wei, F., Huang, S., & Che, W. (2023). Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv*. https://doi.org/10.48550/arXiv.2310.14799

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, *22*, 31–72. https://doi.org/10.1007/s10618-010-0175-9

Silva, V. A., Bittencourt, I. I., & Maldonado, J. C. (2018). Automatic question classifiers: A systematic review. *IEEE Transactions on Learning Technologies*, *12*(4), 485–502. https://doi.org/10.1109/TLT.2018.2878447

Tian, Z., Flanagan, B., Dai, Y., & Ogata, H. (2022). Automated matching of exercises with knowledge components. *30th International Conference on Computers in Education Conference Proceedings*, 24–32.

van Grinsven, M., Brinkhuis, M., Krempl, G., & Snijder, J. (2023). Efficient and general text classification: An active learning approach using active learning and nlp to aid processes such as journalistic investigations and document analysis. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 105–120. https://doi.org/10.1007/978-3-031-74627-7_8

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., & Ma, T. (2023). Larger language models do in-context learning differently. https://doi.org/10.48550/arXiv.2303.03846

Xu, N., Xue, D., Qian, S., Fang, Q., & Hu, J. (2025). Prompting large language models for automatic question tagging. *Machine Intelligence Research*, 1–12. https://doi.org/10.1007/s11633-024-1509-1

Zemlyanskiy, Y., Gandhe, S., He, R., Kanagal, B., Ravula, A., Gottweis, J., Sha, F., & Eckstein, I. (2021). Docent: Learning self-supervised entity representations from large document collections. *arXiv*. https://doi.org/10.48550/arXiv.2102.13247