Journal of Convergence in Technology and Management: Global Nexus
vol. 1, no. 2 (March-April, 2025), pp.1-46

LLM-BASED GENERATION OF EXAMINATION AND PRACTICE TASKS – A SYSTEMATIC LITERATURE ANALYSIS

Niklas Lettow
Chair of Business Administration, esp. Corporate and Management Accounting
FernUniversität in Hagen
Universitätsstraße 41, 58084 Hagen, Germany

Kristopher Pantani
Chair of Business Administration, esp. Corporate and Management Accounting
FernUniversität in Hagen
Universitätsstraße 41, 58084 Hagen, Germany

Prof. Dr. Jörn Littkemann
Chair of Business Administration, esp. Corporate and Management Accounting
FernUniversität in Hagen
Universitätsstraße 41, 58084 Hagen, Germany

ABSTRACT

The increasing role of artificial intelligence in higher education has led to a growing interest in automating the generation of examination and practice tasks. Large language models have emerged as a promising solution to reduce the time and effort required for exam creation while enhancing flexibility and scalability. This systematic literature review provides a comprehensive overview of research on large language model based task generation, examining its effectiveness, methodological approaches

and quality criteria, as well as key challenges. A total of 60 studies were analyzed, covering various domains. The results show that large language models offer enormous potential, especially in terms of time and cost efficiency. While LLM-generated tasks exhibit high grammatical accuracy and contextual relevance, quality varies depending on model fine-tuning, prompt engineering techniques, and training data. Automated evaluation metrics alongside expert and student assessments, reveal that LLM-generated tasks are often indistinguishable from manually created ones. Despite their efficiency, challenges remain, including the risk of hallucinations, bias in generated content, and the lack of standardized evaluation frameworks. Moreover, ethical and legal considerations must be addressed before fully integrating LLMs into exam creation processes.

Keywords: *Task Generation, Automation, Higher Education, Large Language Model, Examinations*

1. INTRODUCTION

At universities, examinations have traditionally been a common practice to assess students' learning progress and acquired subject knowledge. Written examination papers, consisting of one or more test tasks, are a frequently used examination format. For instance, in the Bachelor's program in Economics at the FernUniversität in Hagen, 16 out of a total of 18 required examinations are written exams (FernUniversität in Hagen, 2024). The recurring preparation of examination papers and tasks accordingly demands a significant amount of time from instructors. From the students' perspective, modular programs are often limited by the fact that universities typically offer only one exam date per module per semester. Automating examination processes – including the automated generation of exam tasks and papers as a sub-process – could therefore help to relieve instructors from repetitive routine tasks and provide students with greater flexibility through variable exam schedules. Given limited resources, increasing student numbers, and the central importance of innovation, it is evident that automated task generation will play an increasingly significant role, even in the context of business administration. Initial attempts to automate the creation of exam tasks and papers date back to the 1970s and were practically implemented (Wolfe, 1976). According to Mulla and Gharpure (2023), from a technical perspective, these approaches can be categorized into rule-based methods and those based on artificial neural networks (ANN). However, until the end of the 2010s, the methods used were predominantly rule-based, largely due to technical limitations, particularly limited computing capacity, which made these approaches heavily reliant on the quality of the underlying rules. Despite this, a substantial number of studies also explored ANN-based approaches. Nevertheless, these approaches faced performance constraints, one of the main reasons being that data could only be processed sequentially rather than in parallel (Soni *et al.*, 2019). Moreover, training ANNs was extremely time-intensive, partly for this reason. The *vanishing gradient problem* was another frequently mentioned challenge, where the weights of the training data diminished over time and ultimately disappeared from the ANN's memory (Himaja *et al.*, 2021). These challenges have only been satisfactorily addressed with the advent of generative artificial intelligence (AI). Specifically, the development of the transformer architecture – the foundation of numerous well-known large language models (LLMs) – enabled parallel data processing and resolved the vanishing gradient problem (Vaswani *et al.*, 2017). Prominent examples in this context include BERT (Bidirectional Encoder

Representations from Transformers), GPT (Generative Pre-trained Transformer), and T5 (Text-to-Text Transfer Transformer). Since then, the automated generation of textual content has regained the attention of various stakeholders. In this context, however, it is essential to distinguish between traditional machine learning approaches and newer, LLM-based methods to clearly define the boundaries and potentials of the research field examined in this analysis. Accordingly, the literature now increasingly differentiates between the *pre-ChatGPT era* and the *post-ChatGPT era* (Kıyak, 2023). The goal of this paper is therefore not to cover all ANN-based methods of automated text generation but to provide an overview of the current state of the research field newly established by LLMs. In the context of digital teaching and learning, this primarily includes the LLM-based generation of written exercises and exam tasks. For example, solving a variety of domain-specific tasks is considered an effective method for sustainably anchoring learning content in long-term memory (Lu *et al.*, 2021; Dijkstra, 2022). Access to a comprehensive task repertoire can therefore significantly enhance students' learning success and plays a central role in the learning process (Lu *et al.*, 2021; Wang *et al.*, 2022). However, the manual creation of a large number of high-quality tasks is extremely time-consuming and labor-intensive (Meißner *et al.*, 2024). This also applies to the recurring creation of examination documents and tasks. Since LLMs require only a fraction of the time needed for the manual creation of an equivalent number of tasks, their use in generating exam and practice tasks could offer significant advantages for both students and instructors (Rathi *et al.*, 2024). In addition to time and cost savings, positive aspects such as flexibility, diversity, personalization, sustainability and the promotion of digital skills should be mentioned in this context. It is thus unsurprising that scholarly discourse on this topic has steadily increased since the early 2020s. A comprehensive overview of this emerging research field, however, has not yet been identified. Literature reviews can be found, for example, in Kurdi *et al.* (2020) on the topic of *Automated task generation* or in Artsi *et al.* (2024) on the topic of *LLM-based task generation for medical exams*. However, the first-mentioned article addresses an extremely broad repertoire of methods and, in addition to AI-based tools, also addresses various other instruments. The second-mentioned article, on the other hand, is limited to the use of LLMs, but is narrowly limited in terms of domains, as only the field of medicine is examined. An overview of the research activities on LLM-based task generation with a cross-domain character is still pending. Therefore, the present analysis aims to fill this gap by drawing on previously published research contributions.

This paper adopts an application- and result-oriented perspective, largely omitting technical, ethical, and legal considerations. This does not mean that these aspects are less important, but rather that a comprehensive assessment of their equally importance would go beyond the scope of this article. Investigating into technical, ethical and legal aspects could therefore provide valuable opportunities for future research projects.

To further emphasize the application- and result-oriented perspective pursued here, the paper briefly outlines its objectives. Specifically, the core question is, whether LLMs can generate high-quality exam and practice tasks suitable for use in real exam scenarios. To address this core question, the paper examines a series of subordinate, operationally focused questions:

Which LLM models are used for task generation? What specific types of tasks have been generated? For which application domains have the tasks been generated? Which regions exhibit particular research dynamics, and how are these activities distributed over time? What evaluation methods are employed, and what quality criteria are used to assess the generated tasks? What specific results have been achieved regarding quality criteria, and what patterns can be identified?

The answers to these subordinate questions will ultimately serve to address the overarching core question by providing both a comprehensive overview and a differentiated depiction of the research field under consideration. To this end, the subsequent chapters present a systematic literature analysis. This method was chosen because it enables a comprehensive overview of the entire research field and offers favorable conditions for systematically capturing relevant insights and developing a well-founded understanding of the topic. Chapter 2 explains the methodological approach used. Chapter 3 describes the data collection process, followed by data analysis in terms of presentation, interpretation, and evaluation of the results. Chapter 4, the concluding section, summarizes the key findings and synthesizes them into an overall perspective.

2. METHODOLOGY

The chosen analytical methodology is primarily based on a conceptual proposal by Randolph (2009), which draws upon a taxonomy developed by Cooper. This taxonomy distinguishes between five different analytical characteristics, one of which is the so-called focus-oriented characteristic. This characteristic emphasizes previous research findings, research methods, theories, as well as practices and applications within the

research field under consideration (Cooper, 1988). As the focus-oriented characteristic aligns most closely with the objectives of this analysis, it will be adopted in the subsequent sections of the paper. The analytical process itself, irrespective of the chosen analytical characteristic, is also grounded on a systematic framework developed by Cooper (1984). It encompasses the following steps: *problem formulation* (1), *data collection* (2), *data evaluation* (3), *analysis and interpretation* (4) and *public presentation* (5). The problem formulation (1) has already been addressed in a preliminary manner in the introduction. An overview of the key issues is presented in **Error! Reference source not found.**, juxtaposed with the potential benefits of LLM-based task generation. However, this list may be supplemented with additional points and is therefore not claimed to be exhaustive.

Problems of conventional task generation	Potential of LLM-based task generation
High creation effort	High efficiency (time and cost savings)
Prone to errors (e.g. due to time pressure)	Flexibility (individual examination time freely selectable)
Time inflexibility (only 1 examination date for all)	Diversity (e.g. multilingualism)
Plagiarism protection (risk of copying)	Personalization
Limited variability (repeating identical tasks)	Sustainability (paperless)
…	Promotion of digital skills
	…

Table 1: Problems and Potential

During the next step, data collection (2), scientific databases and catalogs were searched for relevant studies. The literature search was conducted on June 12, 2024. After multiple brainstorming sessions, the keywords shown in Table were selected and combined using logical operators.

<OR>		<OR>		<OR>

gpt*, chatgpt, „generative pretrained transformer*“, llm*, „large language model*, „ki“, „künstliche intelligenz*“, „ai“, „artificial intelligence*“, „machine learning“, „deep learning“, „bert“, „bard“, llama*, claude*, gemini, „palm“, grok, mixtral	<AND>	exam*, assessment*, klausur*, frage*, aufgabe*, prüfung*, question*	<AND>	generat*, generier*, erstell*, erzeug*, education*, higher

Table 2: Connections between Keyword Groups

To avoid an overflow of irrelevant results, the search was limited to titles and abstracts based on the connected keyword groups. Additionally, the search was restricted to German- and English-language results, as the share of non-German and non-English results was only 1.75 %, consisting largely of Chinese-language studies that could not be reliably translated. Books, conference papers, edited volumes, and journal articles were included in the search without any restrictions on the publication period. The data sources included the Bielefeld Academic Search Engine (BASE), the UB catalog/DigiBib of the University Library of the FernUniversität in Hagen, and all databases available through the library in the fields of economics, pedagogy, educational sciences, and computer science. This broad selection was necessary due to the interdisciplinary nature of the research field, which encompasses economic, educational, and technical issues. To exclude low-quality studies, the search was further restricted to peer-reviewed results. Subsequently, the resulting list of studies was screened based on their titles and further narrowed using several context-specific exclusion criteria. In the next step, the abstracts of the remaining studies were reviewed, applying the same exclusion criteria, which led to the elimination of additional results. The remaining studies were then subjected to a full review. During this process, the number of studies included in the analysis was further reduced based on the defined exclusion criteria. Ultimately, all studies meeting one or more of the exclusion criteria listed in **Error! Reference source not found.** were eliminated from the analysis.

Exclusion Criteria
LLMs are not used for (automated) task generation.

The study does not describe a specific application case.
The study discusses the use of LLMs for solving exam tasks.
The study focuses on the use of LLMs for grading and evaluating exam tasks.
The study explores the use of LLMs as a monitoring tool (e.g., detecting academic
The study investigates the use of LLMs for generating medical diagnoses and
The study examines the use of LLMs for generating images or videos.
The study analyzes the use of LLMs for generating financial investment
The study addresses the use of LLMs for generating programming code or metadata.

Table 3: Applied Exclusion Criteria

In summary, all studies not directly related to LLM-based task generation were excluded from full review. Due to the very small number of results specifically addressing the LLM-based generation of exam tasks in the context of higher education, studies related to school education were included. Moreover, studies examining the use of LLMs for generating practice tasks were also incorporated. The inclusion of these two criteria was deemed necessary since only n = 2 studies would have remained otherwise. The inclusion of studies on practice task generation and those from the school context is based on the premise that the tasks are fundamentally subject to the same basic requirements as exam tasks designed for higher education. Both practice and exam tasks aim to assess and foster knowledge and competencies. In both cases, tasks should be clearly formulated, factually accurate, pedagogically sound, and appropriately challenging. The assumption that findings from studies on practice tasks and school-related contexts can be transferred to LLM-based exam task generation in higher education is thus grounded in the substantive and functional similarity of the quality requirements for the tasks. The larger number of included contributions and broader data basis were intended to yield more reliable insights into the variables that significantly influence the quality of LLM-generated tasks. Additionally, a forward and backward search was conducted alongside the full review of studies to identify further relevant results for the analysis. The forward search aimed to identify later works citing a particular source, while the backward search focused on identifying relevant sources cited within a given work.

The results obtained during the data collection (2), as well as the subsequent steps of Cooper's systematic framework – data evaluation (3) and analysis and interpretation (4) – are presented in the following chapter.

3. RESULTS

3.1 Data Collection

The keyword search yielded a total of 3320 results. After removing duplicates as well as studies not published in German or English, 1687 entries remained, which were further screened based on their publication titles. Of the remaining 153 results, the abstracts were reviewed in the subsequent step. Applying the exclusion criteria, the dataset could be reduced to 87 entries. Additionally, 15 further publications were identified through forward and backward citation searches. As a result, the dataset consisted of a total of 102 studies, which were subjected to a full review. During this process, an additional 42 studies were excluded based on the exclusion criteria. Ultimately, 60 contributions remained that were included in the analysis. **Error! Reference source not found.** provides an overview of the entire process.

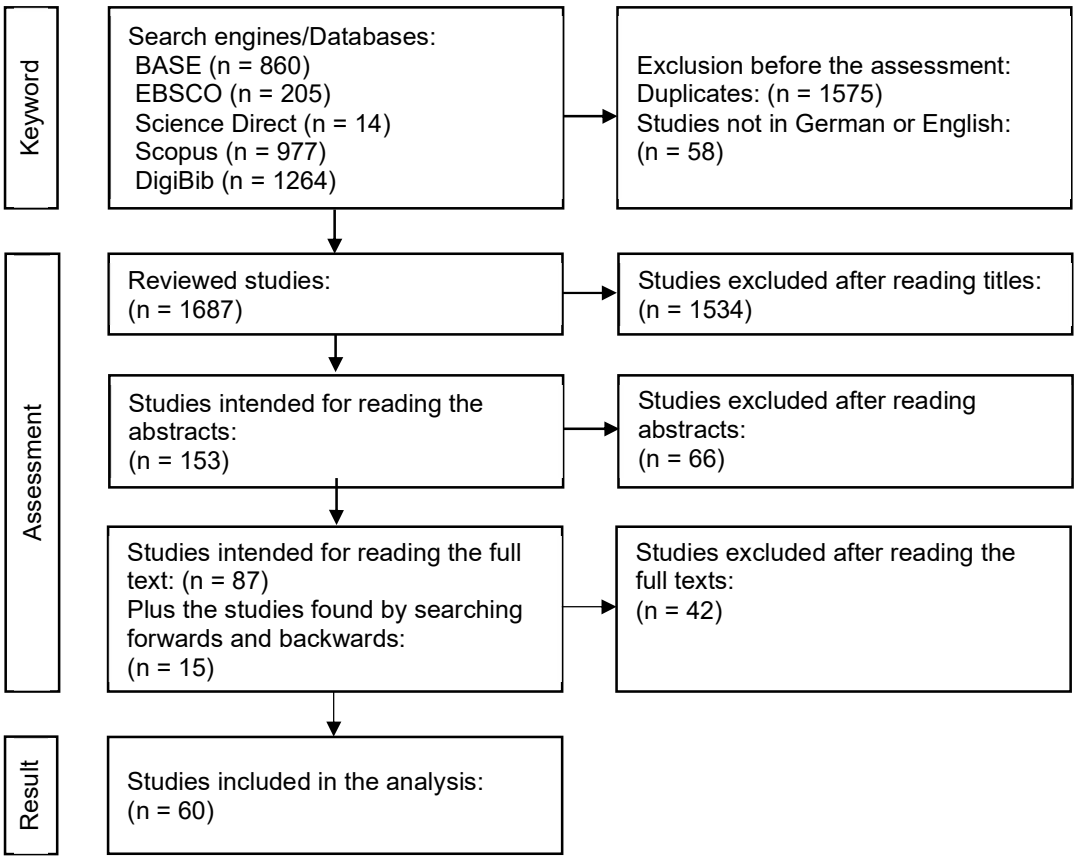

Figure 1: Data Collection Process

3.2 Data Evaluation

1. Which regions show particular research dynamics, and how are these activities distributed over time?

In the context of data analysis (3), it can initially be noted that the majority of research

activities on LLM-based task generation are concentrated in the Asian region, with the United States ranking first at the country level, see Figure 2. It is worth mentioning that all analyzed studies were written in English. However, the LLM-generated questions were not exclusively in English. While English constituted the majority, tasks were also generated in other languages, such as German (Laupichler *et al.*, 2023), French (Bitew *et al.*, 2023; Hudon *et al.*, 2024), Indonesian (Vincentio & Suhartono, 2022; Suhartono *et al.*, 2024) and Swedish (Goran & Abed Bariche, 2023; Kalpakchi & Boye, 2021). Of particular note in this context is a study by Ushio *et al.* (2023), which compares the performance of three different LLMs across eight languages. Chronologically, the two oldest studies included in the analysis are by Chan and Fan (2019), as well as Lopez *et al.* (2020), published in 2019 and 2020, respectively.

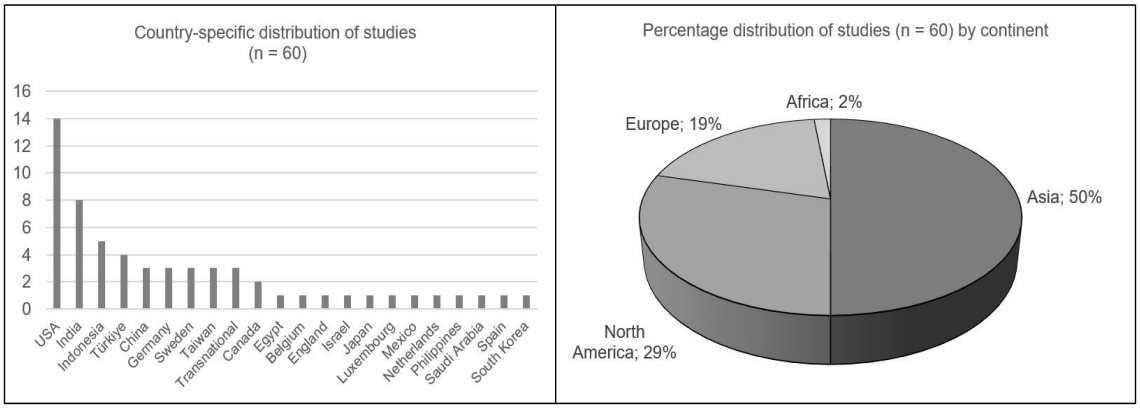

Figure 2: Geographical Distribution of Research Activities

From that point onward, a clear upward trend in the number of publications can be observed, although the number of publications in 2024 shows a decline, see **Error! Reference source not found.**. This, however, is attributable to the fact that the keyword search was conducted on June 12, 2024. Given the ongoing research activities on LLMs, it can be assumed that additional relevant publications have been

released after this date, which could no longer be considered in the present analysis.

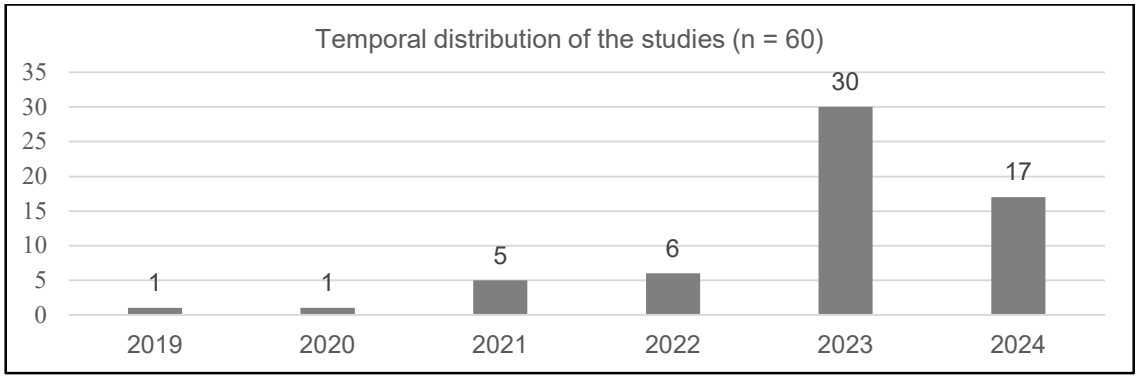

Figure 3: Temporal Distribution of the Studies Included in the Analysis

2. Which LLM models are used to generate tasks?

It can be noted that the GPT models developed by OpenAI account for the largest share by far. Across the 60 analyzed studies, a total of 89 application scenarios were described, with one of the GPT models being used in 55 scenarios (or 61.80 %), as shown in Figure 4. Furthermore, 8 studies employed custom model settings, meaning that multiple LLMs were used within a single application scenario to handle different subtasks. For example, Tsai *et al.* (2021) utilized a BERT model for semantic analysis and keyword extraction from educational materials. Subsequently, they employed Stanford CoreNLP for syntactic analysis and, finally, GPT-2 for question generation. This approach of constructing custom settings was particularly prevalent in 2021 and 2022 but has since declined – presumably due to the increasing performance of off-

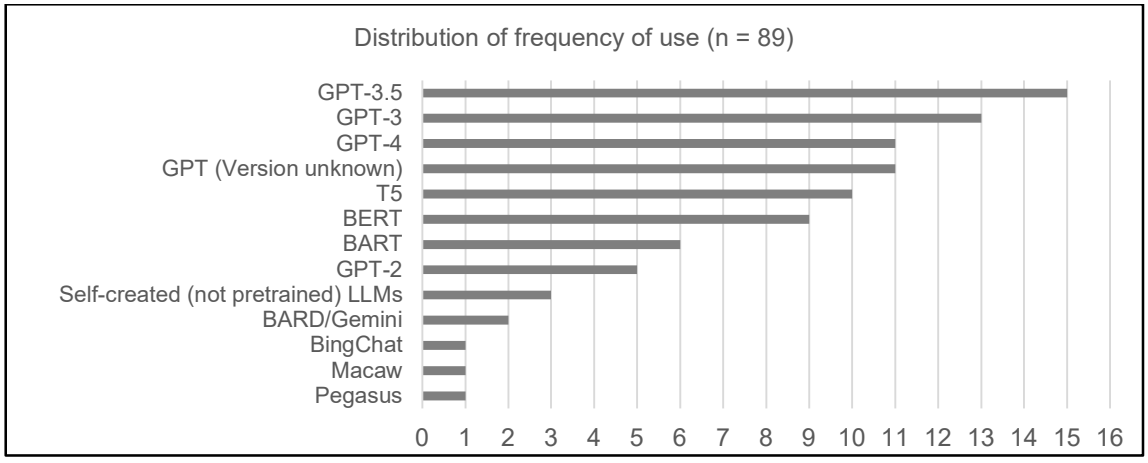

Figure 4: Distribution of Frequency of Use

the-shelf LLMs. As a counterpart to this trend, reference should be made to a study by

Dhanya *et al.* (2022). In their work, they used the aiXAM platform, a holistic system that not only supports LLM-based generation of multiple-choice (MC) questions but also includes features for creating, conducting, and evaluating exams with automated reporting capabilities.

With regard to the operational methodology for task generation, a fundamental distinction can be made between fine-tuning and prompt engineering. Fine-tuning was employed in 26 of the 60 analyzed studies. This approach involves adapting LLMs to specific application contexts through additional training based on specific datasets. Among these, the Stanford Question Answering Dataset (SQuAD) was used in 13 of the 26 studies. SQuAD is a dataset specifically designed for the development and evaluation of LLM-based question-answering systems and contains over 100,000 question-answer pairs derived from Wikipedia articles (Uto *et al.*, 2023). Other datasets frequently used for fine-tuning included RACE (Dijkstra *et al.*, 2023; Rodriguez-Torrealba *et al.*, 2022; Rathi *et al.*, 2024), OpenStax (Wang *et al.*, 2022; Olney, 2023) and TyDiQA (Vincentio & Suhartono, 2022; Suhartono *et al.*, 2024). Additionally, 16 of the 26 studies utilized custom data, such as excerpts from textbooks and educational materials, for fine-tuning. Notably, 10 of these 26 studies employed multiple datasets for fine-tuning to evaluate the performance of LLMs depending on the training data. Moreover, 15 of the 26 studies applied combinations of fine-tuning and prompt engineering. In contrast, no fine-tuning was conducted in 34 of the 60 studies. In these cases, the generation process was achieved solely through prompt engineering. Prompt engineering can be described as the art or technique of designing input prompts to elicit optimal results from the LLM (Elkins *et al.*, 2023). Regarding prompt techniques, it is important to note that they varied significantly between studies, with each being uniquely tailored to the specific use case. However, to achieve at least a broad classification, prompt techniques can be divided into zero-shot, one-shot, and few-shot approaches (Wang *et al.*, 2022; Goran & Abed Bariche, 2023). In summary, one-shot prompting was applied in only three scenarios, while the remaining scenarios were almost evenly distributed between zero-shot and few-shot prompting. A concise overview of the characteristics of these prompt techniques is provided in **Error! Reference source not found.**.

Prompt technique	Characteristics

Zero-Shot-Prompting	The LLM is instructed to generate a task without any examples. It only receives clear, precise instructions.
One-Shot-Prompting	The LLM receives an example of the desired task (e.g. from specific teaching material) and is instructed to generate a task based on this.
Few-Shot-Prompting	The LLM receives several examples of the desired task (e.g. from specific teaching material) and is instructed to generate a task based on these.

Table 4: Zero-Shot-, One-Shot- and Few-Shot-Prompting

3. For which application domains were the tasks generated?

With regard to the application domains in which tasks were generated, the field of medicine, along with the category *General*, clearly ranks first, each accounting for 21.25 % of the total. The *General* category represents studies that did not specify concrete application domains, see **Error! Reference source not found.**. Broadly speaking, the bar chart indicates

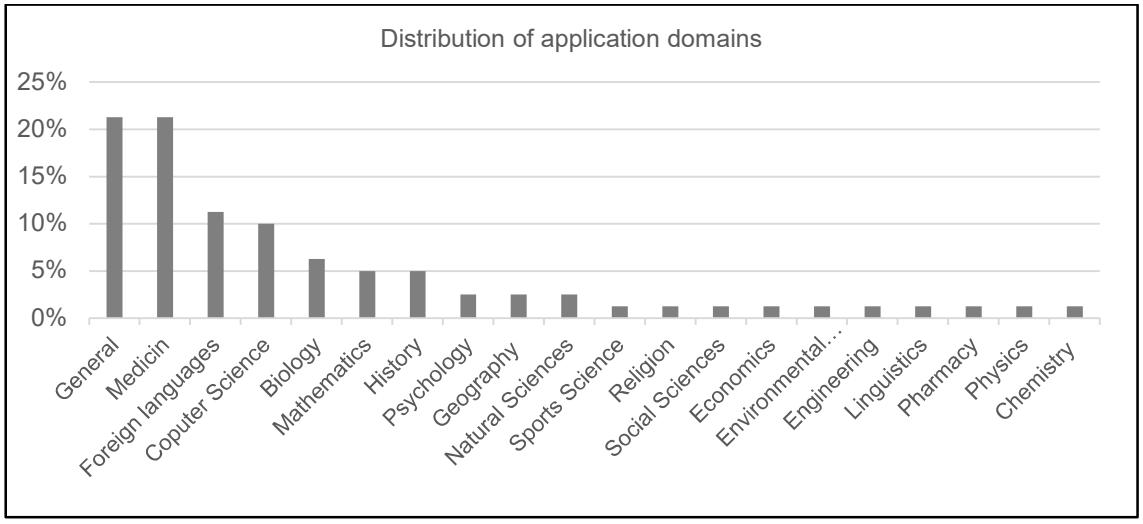

Figure 5: Distribution of Application Domains

that nearly half of all studies are situated in the fields of medicine, foreign languages, computer science, and biology. In contrast, the remaining studies are distributed across various other application domains without any notable concentrations.

4. What specific task types were generated?

As a result, it can be concluded that multiple-choice (MC) tasks, accounting for 48.84 % of the total, represent by far the most frequently observed task type. In this context,

this task type can be considered a singular outlier, whereas the distribution of the remaining task types is more or less balanced, see Figure 6.

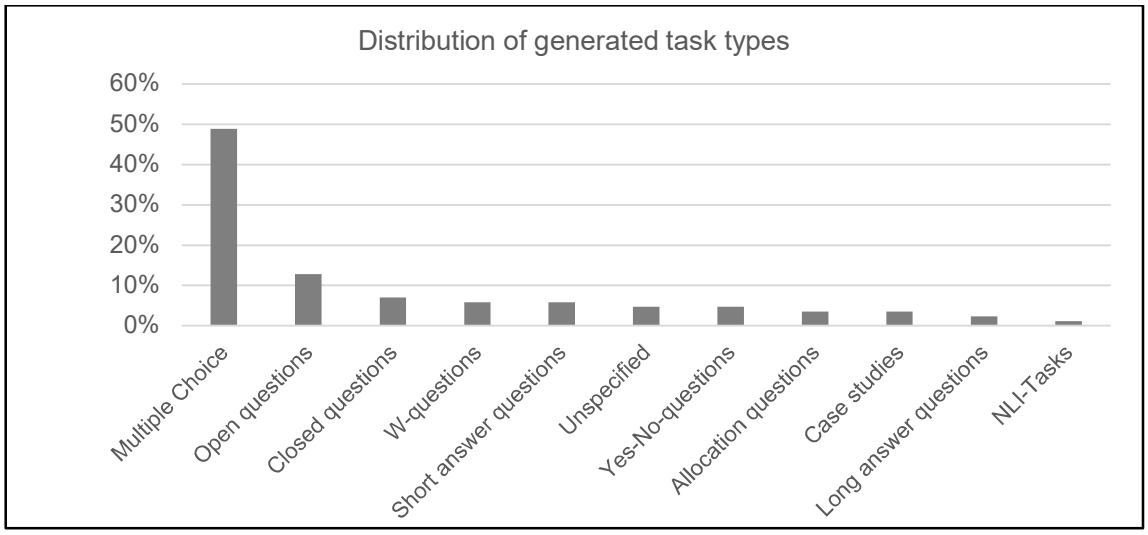

Figure 6: Distribution of Generated Task Types

5. Which evaluation methods are applied, and what quality criteria are used to assess the generated tasks?

The quality assessment of exam tasks fundamentally depends on various factors, encompassing both objective and subjective dimensions. Examiners, learners, and external reviewers often hold differing opinions regarding the requirements a high-quality exam task must meet. Consequently, it is not surprising that a wide range of evaluation methods and quality criteria were identified. Specifically, the data analysis of the 60 studies revealed a total of 59 distinct quality criteria. A tabular overview of which criteria were applied in each study can be found in **Error! Reference source not found.** in the Appendix. It is worth noting that the quality of the generated tasks was assessed in 58 studies using at least two or more criteria. This approach aims to provide a nuanced representation of task quality and to facilitate performance comparisons between LLMs (Ahmed *et al.*, 2024), prompt techniques (Wang *et al.*, 2022), task types (Elshiny & Hamdi, 2023), application domains (Elkind *et al.*, 2023), and fine-tuning datasets (Ushio *et al.*, 2023). Only one of the 60 studies did not specify any explicit quality criteria (Khilnani, 2023). Additionally, another study employed the general suitability of different task types as the sole evaluation criterion (Rai *et al.*, 2023).

The quality assessment was carried out using both automated metrics and subjective evaluation criteria. The automated metrics are quantitative measures collected independently of personal perceptions. In contrast, the subjective evaluation criteria are based on the individual opinions or assessments of evaluators, with the characteristics represented as binary codes, Likert scales, or percentage values. The evaluators consisted of both domain experts, typically educators (Cheung *et al.*, 2023) and members of the target audience (Drori *et al.*, 2022), such as students, for whom the generated tasks were intended. Notably, some scenarios involved testing LLM-generated exam questions on real (often unaware) students, who subsequently evaluated the tasks using a feedback questionnaire following the (simulated) examination (Nasution, 2023). Additionally, combined exam or practice scenarios containing both LLM-generated and manually created tasks were frequently utilized (Drori *et al.*, 2022). This approach is particularly informative when focusing on whether significant differences exist between manually created and LLM-generated tasks. The evaluation process typically begins with defining a set of quality criteria (e.g., grammatical correctness, difficulty level, etc.). This allows raters to assign points for each task based on each quality criterion, for example on a Likert scale. These scores can then be compared across LLM-generated and manually created tasks for each criterion. Distinguishability emerges therefore as an overarching quality criterion, enabling a nuanced understanding of how LLM-generated tasks differ (or do not differ) from manually created tasks across various criteria. In the analyzed studies, distinguishability was assessed using two-sample t-tests (Laupichler *et al.*, 2024) and simple descriptive statistics (Doughty *et al.*, 2024). Another overarching quality criterion frequently employed is inter-rater reliability (IRR), typically measured using Cohen's Kappa (Edwards & Erstad, 2024) or Fleiss' Kappa (Doughty *et al.*, 2024). Like distinguishability, IRR can be determined separately for all human-evaluated quality criteria.

In addition to distinguishability and IRR as overarching quality criteria, the studies commonly employed a mix of automated metrics and subjective criteria. A broad overview reveals that 7 studies relied exclusively on automated metrics, 28 exclusively on subjective criteria, and 24 used a combination of both. The most frequently observed criteria were context relevance, difficulty level, and grammatical correctness

as subjective measures, and the BLEU score (Papineni *et al.*, 2002) and ROUGE-L score (Lin, 2004) as automated metrics.[1]

Context relevance was used in 21 of the 60 studies, making it the most frequently applied quality criterion. It was predominantly measured through human evaluations using Likert scales or binary coding, often reported as the percentage of contextually relevant questions. However, automated metrics like BLEU or ROUGE-L were also applied, despite their limitations in capturing semantic meaning. In some cases, a combination of human evaluations and automated metrics was used to assess context relevance. The second most frequently applied criterion, used in 18 of the 60 studies, was difficulty level. Difficulty was typically evaluated by human experts who assessed the cognitive effort required for each task. Frameworks such as Bloom's taxonomy (Kratwohl, 2002) and Item Response Theory (IRT) (Uto *et al.*, 2023) were commonly employed. Additionally, expert comparisons of LLM-generated and manually created questions were performed (Akbar *et al.*, 2023; Guan *et al.*, 2023). Statistical measures, such as response times and scores per task, were also used to infer difficulty levels (Laverghetta & Licato, 2023). The BLEU score was the third most commonly used metric, observed in 15 of the 60 studies, followed closely by the ROUGE-L score and grammatical correctness, each applied in 14 studies. Grammatical correctness was primarily evaluated using Likert scales, although binary scales (e.g., *grammatically correct* or *grammatically incorrect*) were also employed, with results reported as the percentage of grammatically correct tasks.

Two additional quality criteria of note are the METEOR score (Denkowski & Lavie, 2014) and discrimination index. The METEOR score is often used alongside BLEU and ROUGE-L as an automated metric for measuring semantic similarity. The discrimination index, particularly relevant for multiple-choice tasks, was assessed through correlation analyses, with the point-biserial correlation coefficient being the primary tool (Coşkun *et al.*, 2024).

6. What specific results were achieved with regard to the quality criteria and what connections can be identified?

The evaluation results regarding context relevance were mixed overall, with the majority of studies concluding that LLM-generated exam tasks were factually accurate

[1] A description of the quality criteria highlighted in the article can be found in **Error! Reference source not found.** in the Appendix.

and appropriate for their respective educational contexts. To confirm this general impression, the scores from 12 studies that quantitatively assessed context relevance were normalized to a 0-100 scale using Min-Max normalization (Milligan & Cooper, 1988), where a score of 100 represents the highest possible level of context relevance. The resulting mean score was 81.43, supporting the overall perception of a tendency toward high context relevance, see **Error! Reference source not found.**. However, it should be noted that due to the varying data collection and evaluation methods applied in the individual studies, the aggregated scores are not entirely free from biases and subjective influences and should therefore be interpreted as indicative rather than definitive. Another noteworthy finding, in addition to the generally high context relevance, is that fine-tuned LLMs performed, on average, more than ten points lower than non-fine-tuned models. However, the coefficient of variation was slightly lower for fine-tuned models, indicating that quality differences were generally less

	Mean (Scale 0-100)	**Standard deviation**	**Coefficient of variation**
Context relevance score	**81.43**	**9.20**	**0.1129**
Without Fine-Tuning	86.69	8.52	0.0983
With Fine-Tuning	76.18	6.42	0.0843

Table 5: Aggregated Context Relevance Scores

pronounced compared to models that had not undergone fine-tuning. Furthermore, higher temperature[2] settings tended to result in more creative task designs but were simultaneously associated with a reduction in content quality (Agarwal *et al.*, 2024).

Regarding difficulty level, the quantitative evaluation results from 14 studies were also normalized to a 0-100 scale using Min-Max normalization, where a value of 0 represents extremely low and a value of 100 represents extremely high difficulty, see **Error! Reference source not found.**. The calculated values suggest that, overall, difficulty levels tend to be balanced, with only a relatively small number of tasks classified as extremely easy or extremely difficult.

	Mean (Scale 0-100)	**Standard deviation**	**Coefficient of variation**

[2] Temperature is a parameter to control the creativity and unpredictability of an LLM.

Difficulty score	**52.83**	**23.02**	**0.44**
Without Fine-Tuning	50.41	24.81	0.49
With Fine-Tuning	57.66	17.98	0.31

Table 6: Aggregated Difficulty Score Values

The difficulty scores also incorporated the results from studies that quantified difficulty based on Bloom's Taxonomy (Ahmed *et al.*, 2024; Sihite *et al.*, 2023; Singh *et al.*, 2023). It should be emphasized that, due to methodological constraints, the derived difficulty scores – like the context relevance scores – should be interpreted only as indicative measures.

Regarding grammatical correctness, the results shown in **Error! Reference source not found.** were obtained. These results were also normalized to a 0-100 scale using Min-Max normalization, where a value of 100 represents the highest level of grammatical accuracy. The analysis includes data from 13 studies in which grammatical correctness was quantitatively evaluated. The results indicate that overall grammatical correctness can be classified as high. It was also observed that non-fine-tuned LLMs performed better on average than fine-tuned models.

	Mean (Scale 0-100)	**Standard deviation**	**Coefficient of variation**
Grammar score	**91.62**	**9.34**	**0.10**
Without Fine-Tuning	95.48	5.55	0.06
With Fine-Tuning	89.56	10.25	0.11

Table 7: Aggregated Grammatical Correctness Scores

Furthermore, non-fine-tuned LLMs exhibited lower variability and greater stability in their results, as evidenced by a lower coefficient of variation of 0.06 compared to 0.11 for fine-tuned models. Additionally, Wang *et al.* (2022) found that increasing the number of prompt examples (particularly in five-shot and seven-shot prompting) led to a reduction in the number of grammatical errors.

Regarding the automated metrics BLEU 1-4, ROUGE-L, and METEOR, the descriptive findings shown in Table 8 were obtained.

	BLEU-1 (n = 52)	**BLEU-2 (n = 52)**	**BLEU-3 (n = 52)**	**BLEU-4 (n = 63)**	**ROUGE-L (n = 83)**	**METEOR (n = 43)**

Mean	35.70	18.61	11.84	11.98	39.61	31.39
Median	36.98	16.97	8.93	6.05	41.01	30.70
Minimum	17.13	2.44	0.06	0.01	21.32	14.18
Maximum	55.03	47.81	43.54	61.71	78.41	52.00

Table 8: Descriptive Findings of the Automated Metrics

The sample sizes (n) represent the number of scenarios in which the respective metric was calculated. The BLEU-1 score indicates that, on average, 35.70 % of unigram pairs (1-grams) between the generated tasks and the reference texts matched. However, as the n-gram size increases (BLEU-2 to BLEU-4), the average scores tend to decline, suggesting that longer n-grams were less frequently observed. For example, the minimum value for the BLEU-4 metric is close to zero, indicating cases where the generated exam tasks shared no longer n- grams with the references. From the BLEU-4 metric's mean score of 11.98 and its median of 6.05, it can be inferred that there is generally low alignment but high creativity in text generation. This finding is further supported by the ROUGE-L metric's mean score of 39.61. While this score is significantly higher than the BLEU-4 mean, it is still considered moderate, as the ROUGE-L metric is less sensitive to minor deviations and more tolerant of synonyms and alternative phrasings compared to BLEU-4. When the METEOR score is additionally considered, further insights into the semantic quality of the generated exam tasks can be derived. Specifically, the mean METEOR score of 31.39 indicates a generally moderate level of semantic alignment between the generated tasks and the reference texts. However, the range of 14.18 to 52.00 reveals a relatively high degree of variability in the data. A correlation analysis between the METEOR metric and the other metrics, see **Error! Reference source not found.**, provides additional context for interpreting these findings. The correlation coefficients between METEOR and BLEU-1 through BLEU-3

	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE-L**
METEOR	0.9208	0.9702	0.9740	0.5373	0.2015

Table 9: Pearson Correlations between the Automated Metrics

indicate a strong positive relationship between semantic quality and the proportions of 1-grams, 2-grams, and 3-grams. This suggests that a higher proportion of these shorter n-grams is generally associated with higher semantic quality in the generated tasks. In contrast, the relationship between METEOR and BLEU-4 is only moderate, and

between METEOR and ROUGE-L it is weakly positive. Since BLEU-4 and ROUGE-L place greater emphasis on longer and more precise matches, these findings suggest that BLEU-4 and ROUGE-L play a relatively minor role in assessing semantic quality. This leads to the hypothesis that achieving semantic or content alignment is less dependent on matches of longer text sequences, such as 4-grams. Instead, the presence of matches for individual keywords and short sequences between the generated tasks and the reference texts appears to be significantly more important.

With regard to the multilingual capabilities of LLMs, a study by Ushio *et al.* (2023) is particularly noteworthy. In their study, tasks were generated in eight different languages using three distinct LLMs: mT5SMALL, mT5BASE, and mBART. Across languages, all three LLMs produced broadly similar results in terms of BLEU-4, ROUGE-L, and METEOR scores. Notably, mT5BASE performed marginally better than mBART despite having significantly fewer parameters, see Figure 7.

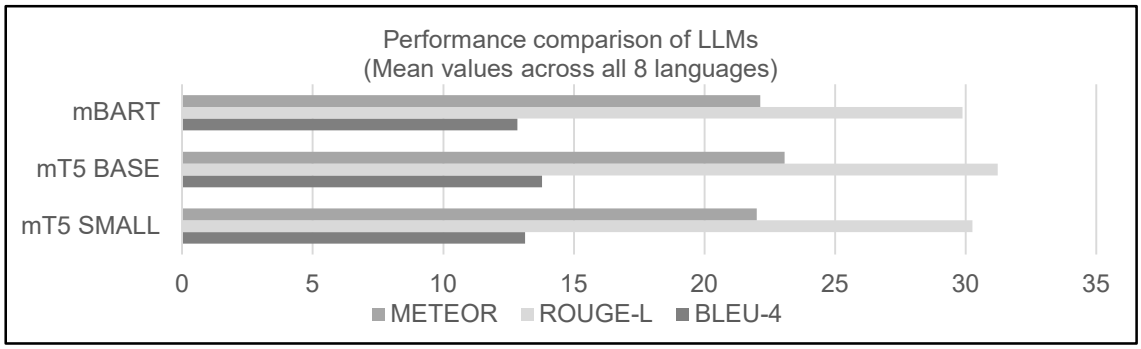

Figure 7: Performance Comparison of LLMs

Across LLMs, the values of the automated metrics revealed significant language-specific differences, as shown in Figure 8. To aid understanding, it should be noted that the language-separated column blocks refer to the study by Ushio *et al.* (2023), while the rightmost outlined column block represents the average values from all analyzed studies. The data indicate that the three LLMs in the study performed relatively well in English, Russian, Japanese, and, in terms of the METEOR score, Korean. In contrast, poorer results were observed for Spanish, Italian, and French, with

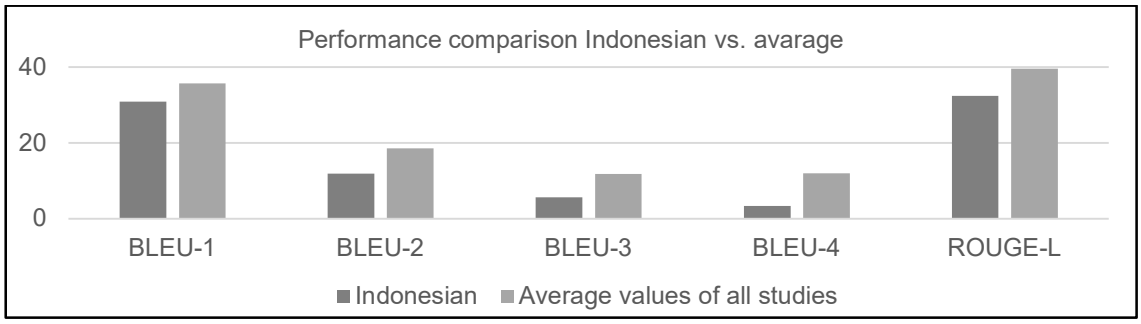

Figure 8: Performance Comparison Indonesian vs. Average

German showing the lowest performance overall. A similar comparison can be drawn from two additional studies that generated tasks in Indonesian (Vincentio & Suhartono, 2022; Suhartono *et al.*, 2024). Both studies applied BLEU 1-4 and ROUGE-L metrics, revealing that tasks generated in Indonesian also showed room for improvement on average, see Figure 9.

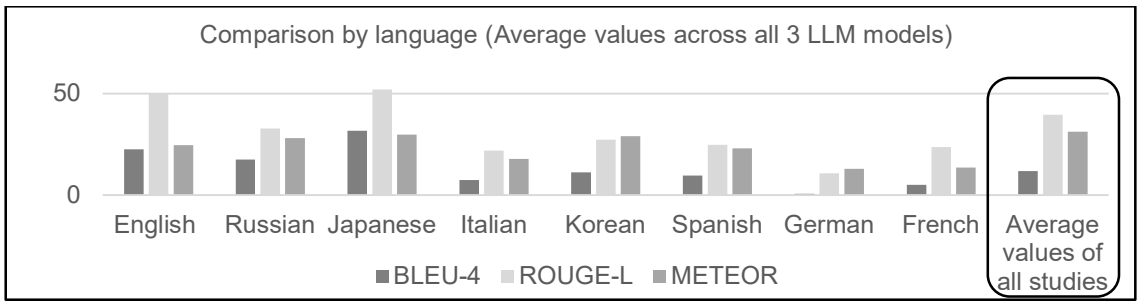

Figure 9: Language-Specific Comparison

Another relationship identified in the analysis of automated metrics is that few-shot prompts are generally associated with higher METEOR scores compared to one-shot or zero-shot prompting techniques. An illustrative finding is presented in the study by Wang *et al.* (2022), see Figure 10. Their study shows that few-shot prompting techniques (e.g., five-shot and seven-shot as shown in the figure) result in a higher proportion of acceptable tasks than zero-shot and one-shot techniques.

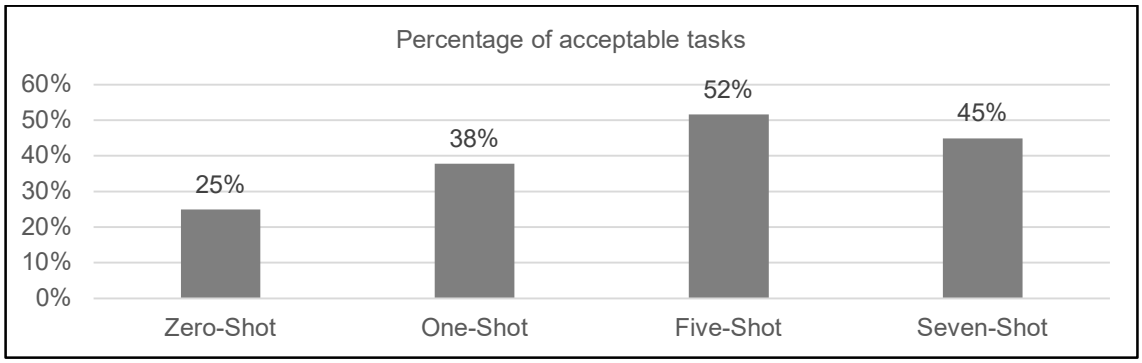

Figure 10: Proportions of Acceptable Tasks depending on the Prompting Technique

Paradoxically, the proportion of acc eptable tasks decreases again in the seven-shot scenario. Regarding the two overarching quality criteria, distinguishability and IRR, the following observations can be made: For distinguishability, the overall results indicate that in 4 out of 12 studies investigating this question, LLM-generated tasks significantly differed from conventionally created tasks (Maity *et al.*, 2023; Olney, 2023; Xiao, 2023; Doughty *et al.*, 2024). However, in 8 studies, raters were unable to clearly classify tasks as either LLM-generated or conventionally created (Lu *et al.*, 2021; Drori *et al.*, 2022; Wang *et al.*, 2022; Fleming *et al.*, 2023; Cheung *et al.*, 2023; Coşkun *et al.*, 2024; Hudon, 2024; Laupichler *et al.*, 2024). In some cases, a more nuanced picture emerged. For example, Olney (2023) found that for 6 out of 7 quality criteria, LLM-generated tasks were indistinguishable from conventionally created tasks. However, in terms of overall quality, conventionally created tasks were rated significantly higher. Similarly, Laupichler *et al.* (2024) reported no significant difference in difficulty level between the two types of tasks, though the discrimination index of LLM-generated tasks was substantially higher than that of conventionally created tasks. The higher discrimination index was frequently attributed to the fact that distractors in LLM-generated multiple-choice (MC) tasks were often not of sufficient quality to effectively differentiate between high-performing and low-performing test-takers. Nevertheless, research specifically addressing the generation of high-quality distractors already exists and could help address this issue (Chung *et al.*, 2020; Offerijns *et al.*, 2020; Feng *et al.*, 2024; Lee *et al.*, 2024).

Regarding IRR, it was observed that the quality assessments among evaluators varied considerably, ranging from strong agreement to significant divergence, see Table 10.

	Cohen´s Kappa	Fleiss´ Kappa	Gwet´s AC 1	Goodman Kruskal´s Gamma
Mean	0.58	0.63	0.87	0.85
Median	0.60	0.70	0.90	-
Minimum	0.10	0.07	0.62	-
Maximum	0.85	0.98	0.96	-

Table 10: Inter-Rater Reliability Results

Specifically, Cohen's Kappa values ranged from 0.1 (Edwards and erstad 2024) to 0.85 (Agarwal *et al.* 2023) (n = 5). The highest values were achieved by Agarwal *et al.* (2023) for multiple-choice (MC) tasks in the medical domain, particularly concerning the quality criteria of *required cognitive effort*, *difficulty level*, and *validity*. For Fleiss' Kappa, the range was even broader, with values between 0.07 (Doughty *et al.*, 2024) and 0.98 (Olney, 2023) (n = 5). The study by Olney (2023) achieved the highest values in this category. In Olney's work, MC questions were also generated for the medical domain, and the IRR was calculated for a total of six different quality criteria. Overall, despite the notable variability, there is a clear tendency for IRR to be classified as moderate to high, as evidenced by median values of 0.6 for Cohen's Kappa and 0.7 for Fleiss' Kappa. Additionally, two other studies are worth mentioning. The first, conducted by Doughty *et al.* (2024), used Gwet's AC1 score to measure IRR. Applied to six quality criteria, the values ranged from 0.62 to 0.96, with a median of 0.90, indicating near-perfect agreement among the raters. The second study, by Kalpakchi & Boye (2021), employed Goodman-Kruskal's Gamma to determine IRR, achieving a value of 0.85 across all quality criteria, which also reflects very high agreement among raters. It is important to note that the aggregated overall IRR values should be interpreted only as rough indicators due to the wide range of data variability. Even within individual studies, significant fluctuations were observed. For example, in the study by Bitew *et al.* (2023), Cohen's Kappa for tasks in the English domain was 0.29, while tasks in the geography domain achieved a value of 0.52.

Regarding time and cost efficiency, a relatively consistent overall picture emerges: LLMs enable the generation of exam and practice tasks in significantly less time compared to the labor-intensive manual creation process by human experts (Johnson

et al., 2023; Kıyak, 2023; Klang *et al.*, 2023; Goyal *et al.*, 2024; Meißner *et al.*, 2024). Rivera-Rosas *et al.* (2024) reported a task generation time of 10-15 seconds per task. Furthermore, Goran & Abed Bariche (2023) estimated the cost of generating 10 tasks at $0.0824, further underscoring the efficiency and potential applications of LLMs in the educational sector.

To provide a final overall impression and address the central question of whether LLMs can generate high-quality exam and practice tasks suitable for use in real examination scenarios, the following can be concluded: Based on 12 studies (20.00 %), the generated tasks can be classified as usable. However, in 47 studies (78.33 %), the tasks were deemed only conditionally usable, meaning that modifications or manual revisions would be necessary before implementation. Additionally, one study (1.67 %) concluded that LLMs are not yet sufficiently advanced to meet the requirements for use in real-world teaching, learning, and examination scenarios (Ayub *et al.*, 2023). An overview of the respective overall assessments derived from the individual studies is provided in **Error! Reference source not found.**. These evaluations were drawn either explicitly or implicitly from the conclusions of the studies. It should be noted that these qualitative assessments are subject to subjective influences and potential biases, which limit their overall interpretive reliability.

#	Article	Results/tasks usable?	#	Article	Results/tasks usable?	#	Article	Results/tasks usable?
1	Agarwal/Sharma/ Goswami (2023)	conditionally	21	Grover *et al.* (2021)	yes	41	Nasution (2023)	conditionally
2	Agarwal *et al.* (2024)	conditionally	22	Guan *et al.* (2023)	conditionally	42	Ngo *et al.* (2024)	conditionally
3	Ahmed et al (2024)	conditionally	23	Himaja *et al.* (2021)	conditionally	43	Olney (2023)	conditionally
4	Akbar *et al.* (2024)	conditionally	24	Hudon *et al.* (2024)	conditionally	44	Onal/Kulavuz-Onal (2023)	conditionally
5	Ayub *et al.* (2023)	no	25	Johnson *et al.* (2023)	yes	45	Rai/Deng/ Liu (2023)	conditionally
6	Bitew *et al.* (2023)	conditionally	26	Kalpakchi/Boye (2021)	conditionally	46	Rathi *et al.* (2024)	conditionally

7	Chan/Fan (2019)	yes	27	Khilnani (2023)	conditionally	47	Rivera-Rosas *et al.* (2024)	conditionally
8	Cheung *et al.* (2023)	yes	28	Kic-Drgas/Kılıçkaya (2024)	conditionally	48	Rodriguez-Torrealba/Garcia-Lopez/Garcia-Cabot (2022)	conditionally
9	Coşkun/Kıyak/Budakoğlu (2024)	conditionally	29	Kıyak (2023)	conditionally	49	Sihite/Meisuri/Sibarani (2023)	conditionally
10	Dhanya/Balji/Akash (2022)	conditionally	30	Kıyak *et al.* (2024)	conditionally	50	Singh/Patvardhan/Vasantha Lakshmi (2023)	conditionally
11	Dijkstra *et al.* (2022)	conditionally	31	Klang *et al.* (2023)	conditionally	51	Stadler/Horrer/Fischer (2024)	conditionally
12	Doughty *et al.* (2024)	conditionally	32	Laupichler *et al.* (2024)	conditionally	52	Suhartono/Majiid/Fredyan (2024)	conditionally
13	Drori *et al.* (2022)	yes	33	Laverghetta/Licato (2023)	conditionally	53	Tran *et al.* (2023)	conditionally
14	Edwards/Erstad (2024)	conditionally	34	Lee *et al.* (2023)	conditionally	54	Tsai/Chang/Yang (2024)	yes
15	Elkins *et al.* (2023)	yes	35	Lee *et al.* (2024)	conditionally	55	Ushio/Alva-	conditionally

							Manchego /Camacho -Collados (2023)	
16	Elshiny/Hamdi (2023)	conditionally	36	Liang *et al.* (2023)	yes	56	Uto/Tomikawa/Suzuki (2023)	conditionally
17	Fleming *et al.* (2023)	conditionally	37	Lopez *et al.* (2020)	conditionally	57	Vincentio/ Suhartono (2022)	conditionally
18	Goran/Abed Bariche (2023)	conditionally	38	Lu *et al.* (2021)	yes	58	Wang *et al.* (2022)	conditionally
19	Goyal/Kumar/Singh (2024)	yes	39	Maity/Deroy /Sarkar (2023)	conditionally	59	Wu *et al.* (2023)	yes
20	Grévisse (2023)	conditionally	40	Meißner *et al.* (2024)	conditionally	60	Xiao *et al.* (2023)	yes

Table 11: Overall Assessment of the Usability of LLM-generated Tasks

3.3 Analysis and Interpretation

As the results demonstrate, LLMs offer substantial potential in the field of automated generation of practice and exam tasks. Each reviewed study adopts a unique operational approach, whether in terms of prompt techniques, task type selection, decisions regarding fine-tuning, or performance measurement. Regarding the objectives and research questions pursued in the studies, it is evident that the primary focus lies on increasing efficiency and alleviating teachers' workload for routine tasks. Nearly all contributions included in the analysis emphasize the significant time investment required for the manual creation of high-quality practice and exam tasks. Some studies also aim to draw comparisons between human-created and LLM-generated tasks. However, it becomes apparent that the field of research is still in its early stages, as only two studies reported the use of LLM-generated tasks in real examination scenarios (Kıyak *et al.*, 2024; Rivera-Rosas *et al.*, 2024). An additional 10 studies (Coşkun *et al.*, 2024; Kalpakchi & Boye, 2021; Kic-Drgas & Kilickaya, 2024; Laupichler *et al.*, 2024; Laverghetta & Licato, 2023; Lu *et al.*, 2021; Nasution, 2023; Sihite *et al.*, 2023; Tsai *et al.*, 2024; Xiao *et al.*, 2023) utilized LLM-generated tasks in

simulated exam or practice scenarios, highlighting that 80 % of the reviewed studies did not involve practical implementation. Another key finding is that, in most observed use cases, only English-language tasks were generated. The use of LLMs for generating practice and exam tasks in other languages remains relatively underexplored. Furthermore, the majority of research contributions have focused predominantly on the generation of MC tasks. Incorporating other task types, such as open-ended questions or case studies, presents significant research potential to further evaluate and optimize the applicability of LLMs in exam generation in the future. Regarding performance, no consistent pattern emerged across all analyzed studies, as it is not possible to definitively identify which specific LLM delivers the best overall results. This is partly due to the multitude of input variables (e.g., LLM type, task type, application domain, fine-tuning (yes/no), prompt technique, etc.) underlying each use case and partly due to the performance measurement itself, which can be conducted using various quality criteria. However, GPT-based models tend to stand out for their versatility and adaptability, while BERT is particularly valued when combined with domain-specific adaptations (Suhartono *et al.*, 2024; Onal & Kulavuz Onal, 2023). Nevertheless, the quality criteria selected for performance measurement are not always clearly delineated and often include subjective components, which can introduce a degree of uncertainty. This uncertainty raises potential conflicts in validity and reliability. Combined with the high number of degrees of freedom, direct performance comparisons between LLMs are only meaningful under certain conditions, as the quality criteria and evaluation methods vary significantly across studies. This challenge of limited objective comparability represents a significant limitation of the present analysis. To mitigate biases stemming from subjective evaluations, the use of automated metrics offers a viable approach. Strong positive correlations were observed between certain automated metrics, particularly between the BLEU-1 to BLEU-3 scores and the METEOR score, which serve as measures of semantic quality. Grover *et al.* (2021) however point out that contentually accurate and contextually relevant questions can also align with low BLEU scores, particularly when synonyms or alternative sentence structures are used. Although automated metrics represent objective criteria, it becomes evident that the interpretation of their specific values is ultimately itself subject to subjective influences. Therefore, the values of automated metrics should always be considered in conjunction with those of other quality criteria (Rodriguez-Torrealba *et al.*, 2022). In the future, standardized

evaluation frameworks incorporating a set of unified assessment methods and selected quality criteria could help improve overall comparability. A preliminary approach to this has been proposed by Ushio *et al.* (2023). Another unresolved question, which remains unclear due to limited comparability, is whether fine-tuning is strictly necessary to achieve reliable and usable results. It is undisputed that the more specific the application context, the more precisely LLMs must be adapted to that context to produce usable results. This is particularly important when generating exam tasks, which cannot be designed arbitrarily but must be precisely aligned with specific teaching content to ensure high context relevance. This can be achieved through targeted prompt engineering, fine-tuning of the models, or a combination of both approaches. The hypothesis that fine-tuning always improves result quality (Tsai *et al.*, 2021) could neither be confirmed nor refuted based on the data in this analysis. Surprisingly, scenarios where LLMs were not fine-tuned performed better in terms of context relevance and grammatical correctness. It could not be substantiated that the quality of LLM-generated tasks improves over the course of the study period, despite the rapid development of LLMs. Furthermore, no clear tendency could be identified as to which task types are particularly suitable for LLM-based generation and which are not. However, these results should be interpreted with caution due to the described measurement conditions and the small sample size, as several studies have also yielded contradictory findings. Generalizing the results to the broader population should therefore always be approached with restraint and with consideration of the underlying methodological limitations. As observed, targeted prompt engineering can already produce high-quality results (Liang *et al.*, 2023). However, fine-tuning may offer additional long-term benefits, particularly when the goal is to generate large volumes of tasks. This could be relevant, for example, in conducting location- and time-independent individual examinations, which could be generated by the examinees themselves at the push of a button, without the need for complex prompts. The significantly higher labor and financial costs associated with fine-tuning compared to prompt engineering must be weighed individually depending on the intended application. However, findings already exist demonstrating that high-quality results can be achieved with relatively few training data and training epochs (Wu *et al.*, 2023). It can be stated with certainty that the quality of generated questions is highly dependent on the quality of the training datasets (Suhartono *et al.*, 2024). This conclusion is also supported by the results of several studies, such as Wu *et al.* (2023) or Singh *et al.*

(2023). Furthermore, the balanced difficulty level observed across all studies is an indicator of LLMs' potential suitability, assuming that examinations designed for comprehensive performance evaluation should include both challenging and less complex tasks.

One problem that still exists, however, is the risk of hallucinating (Elkins *et al.*, 2023; Lee *et al.*, 2023; Kıyak *et al.*, 2024). Grévisse (2023), for example, emphasizes that implausible distractors and grammatical errors can reduce the difficulty level of tasks. This allows examinees to infer the correct answer based solely on grammar, such as when a singular term is sought in the question text, but all incorrect answer options (distractors) are written in the plural. Consequently, it has been repeatedly stressed that a careful manual review of LLM-generated tasks remains essential before these can be used in real practice and examination scenarios (Khilnani, 2023; Klang *et al.*, 2023; Lee *et al.*, 2023; Nasution, 2023; Onal & Kulavuz-Onal, 2023; Tran *et al.*, 2023; Edwards & Erstad, 2024; Kic-Drgas & Kilickaya, 2024; Ngo, 2024; Rivera-Rosas, 2024; Stadler *et al.*, 2024). However, it is reasonable to assume that the frequency of hallucinations will decrease over time due to ongoing research and development in the field of LLMs, as the models become increasingly powerful and reliable (Coşkun *et al.*, 2024). But how reliable does an LLM ultimately have to be to make human supervision completely unnecessary? And even if the *hallucination-free LLM* existed, another question would be to what extent AI-based exam creation without any human review is legally compliant. Should such a practice fall under the umbrella of freedom of research and teaching or not? In what way can an efficient review process be implemented if each examinee is assessed using an exclusive, individually generated LLM-based exam? These and other fundamental questions must be addressed before the fully automated use of LLMs in exam creation can be considered not only technically feasible but also ethically, legally, and pedagogically responsible.

4. SUMMARY

The aim of this literature review was to systematically capture and evaluate the current state of research on LLM-based generation of exam and practice tasks. The central question focused on whether LLMs can generate high-quality tasks suitable for use in real examination scenarios. To this end, 60 studies covering a wide range of application fields and methodological approaches were analyzed. The findings highlight that LLMs, particularly due to their efficiency advantages such as time and

cost savings, represent a promising tool for both educators and students. LLMs enable flexible and highly scalable task creation, characterized by strong contextual relevance, grammatically accurate formulations, and balanced difficulty levels. However, the studies also point to key challenges: the results indicate that human review remains essential before LLM-generated tasks can be responsibly used in real examination scenarios. Furthermore, the quality of the tasks heavily depends on the structuring of prompts and the quality of training data. While fine-tuning can be tailored to meet specific requirements, prompt engineering offers a more pragmatic solution but remains limited in terms of automation. Another outcome of the analysis is the methodological heterogeneity of the studies, which complicates comparability. To derive general trends despite the diverse quality criteria employed, the five most commonly used criteria, as well as two overarching criteria, were analyzed in detail. Several notable relationships were identified and presented. Automated evaluation methods, such as BLEU and ROUGE-L metrics, were frequently employed but provide only partial insights into the semantic quality of tasks. Moreover, the practical use of LLM-generated tasks in real examination scenarios remains rare, as most studies focus on simulation-based or purely conceptual investigations. The implications of these findings underscore the need for further research, particularly in the standardization of evaluation methods and the optimization of models for specific application domains in different languages. Additionally, ethical and legal questions must be addressed to ensure the responsible and practical use of LLMs in the context of university examinations in the long term.

REFERENCES

Agarwal, M., Sharma, P., & Goswami, A. (2023). Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. *Cureus 15(6).* https://doi.org/10.7759/cureus.40977

Agarwal, A., Mittal, K., Doyle, A., Sridhar, P., Wan, Z., Doughty, J. A., Savelka, J., & Sakr, M. (2024). Understanding the Role of Temperature in Diverse Question Generation by GPT-4. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2, 1550-1551*. https://doi.org/10.1145/3626253.3635608

Ahmed, W. M., Azhari, A. A., Alfaraj, A., Alhamadani, A., Zhang, M., & Lu., C. T. (2024). The Quality of AI-Generated Dental Caries Multiple Choice Questions: A Comparative Analysis of ChatGPT and Google Bard Language Models. *Helyon 10*. https://doi.org/10.1016/j.heliyon.2024.e28198

Akbar, M. F., Al Faraby, S., Romadhony, A., & Adiwijaya, A. (2023). Multimodal Question Generation using Multimodal Adaptation Gate (MAG) and BERT-based Model. *In IEEE 8th International Conference for Convergence in Technology (pp. 1-7).*

Artsi, Y., Sorin, V., Konen, E., Glicksberg, B. S., Nadkarni, G., & Klang, E. (2024). Large language models for generating medical examinations: systematic review. *BMC Medical Education 24*, 354. https://doi.org/10.1186/s12909-024-05239-y

Ayub, I., Hamann, D., Hamann, C. R., & Davis, M. J. (2023). Exploring the Potential and Limitations of Chat Generative Pre-trained Transformer (ChatGPT) in Generating Board-Style Dermatology Questions: A Qualitative Analysis. *Cureus 15(8)*. https://doi.org/10.7759/cureus.43717

Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023). Distractor generation for multiple-choice questions with predictive prompting and large language models. *In The 1st International Tutorial and Workshop on Responsible Knowledge Discovery in Education*. https://doi.org/10.48550/arXiv.2307.16338

Chan, Y. H., & Fan, Y. C. (2019). A Recurrent BERT-based Model for Question Generation. *In Proceedings of the Second Workshop on Machine Reading for Question Answering (pp. 154-162)*. https://doi.org/10.18653/v1/D19-5821

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions - A multinational prospective study. *PLoS ONE 18(8)*. https://doi.org/10.1371/journal.pone.0290691

Chung, H. L., Chan, Y. H., & Fan, Y. C. (2020). A BERT-based Distractor Generation Scheme with Multi-tasking and Negative Answer Training Strategies. *Findings of the Association for Computational Linguistics: EMNLP 2020, 4390-4400*. https://doi.org/10.18653/v1/2020.findings-emnlp.393

Cooper, H. M. (1984). The integrative research review: A systematic approach. *In Applied social research methods series 2.*

Cooper, H. M. (1988). Organizing knowledge synthesis: A taxonomy of literature reviews. *Knowledge in Society 1, 104-126.*

Coşkun, Ö., Kıyak, Y. S., & Budakoğlu, I. İ. (2024). ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. *Medical Teacher, 1-7.* https://doi.org/10.1080/0142159X.2024.2327477

Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. *In Proceedings of the ninth workshop on statistical machine translation (pp. 376-380).* https://doi.org/10.3115/v1/W14-3348

Dhanya, N. M., Balaji, R. K., & Akash, S. (2022). AiXAM - AI assisted Online MCQ Generation Platform using Google T5 and Sense2Vec. *In Proceedings of the 2nd International Conference on Artificial Intelligence and Smart Energy (pp.* 38-44).

Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). Reading Comprehension Quiz Generation using Generative Pre-trained Transformers. *In 23d International Conference on Artificial Intelligence in Education (pp. 4-17).*

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2024). A Comparative Study of AI-Generated (GPT-4) and Human-crafted MCQs in Programming Education. *In ACM International Conference Proceeding Series (pp. 114-123).* https://doi.org/10.1145/3636243.3636256

Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., Liu, K., Chen, L., Tran, S., Cheng, N., Wang, R., Singh, N., Patti, T.L., Lynch, J., Shporer, A., Verma, N., Wu, E., & Strang, G. (2022). A Neural Network Solves, Explains, and Generates University Math Problems by Program Synthesis and Few-Shot Learning at Human Level. *Proceedings of the National Academy of Sciences of the United States of America 119(32)*. https://doi.org/10.1073/pnas.2123433119

Edwards, C. J., & Erstad, B. L. (2024). Evaluation of a Generative Language Model Tool for Writing Examination. *American Journal of Pharmaceutical Education 88.* https://doi.org/10.1016/j.ajpe.2024.100684

Elkins, S., Kochmar, E., Serban, I., & Cheung, J. C. K. (2023). How Useful are Educational Questions Generated by Large Language Models? *AIED Late Breaking Results, 536-542*. https://doi.org/10.1007/978-3-031-36336-8_83

Elshiny, R. M., & Hamdy, A. (2023). Automatic Question Generation Using Natural Language Processing and Transformers. *In 5th International Conference on Computer and Applications.*

Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., Ornelas, N. O., & Lan, A. (2024). Exploring Automated Distractor Generation for Math Multiple-choice Questions via Large Language Models. *Findings of the Association for Computational Linguistics, 3067-3082.* https://doi.org/10.18653/v1/2024.findings-naacl.193

FernUniversität in Hagen (2024). Studien- und Prüfungsinformationen Wintersemeseter 2024/2025 der Fakultät für Wirtschaftswissenschaft. https://moodle.fernuni-hagen.de/mod/resource/view.php?id=128401

Fleming, S. L., Morse, K., Kumar, A., Chiang, C. C., Patel, B., Brunskill, E., & Shah, N. (2023): Assessing the Potential of USMLE-Like Exam Questions Generated by GPT-4. Cold Spring Harbor Laboratory.

Goran, R., & Abed Bariche, D. (2023). Leveraging GPT-3 as a question generator in Swedish for High School teachers.

Goyal, R., Kumar, P., & Singh, V. P. (2024). Automated Question and Answer Generation from Texts using Text-to-Text Transformers. *Arabian Journal for Science and Engineering 49, 3027-3041.* https://doi.org/10.1007/s13369-023-07840-7

Grévisse, C. (2023). Comparative Quality Analysis of GPT-Based Multiple Choice Question Generation. *In International Conference on Applied Informatics (pp.* 435-447). https://doi.org/10.1007/978-3-031-46813-1_29

Grover, K., Kaur, K., Tiwari, K., Kumar, R., & Kumar, P. (2021). Deep Learning Based Question Generation Using T5 Transformer. *Communications in Computer and Information Science 1367, 243-255.* https://doi.org/10.1007/978-981-16-0401-0_18

Guan, M., Mondal, S. K., Dai, H. N., & Bao, H. (2023). Reinforcement learning-driven deep question generation with rich semantics. *Information Processing & Management 60.* https://doi.org/10.1016/j.ipm.2022.103232

Himaja, G., Gadu, S. H., Harshith, K. V., Yamini, M., Sravya, S. S., Murahari, K. V., & Mannava, S. (2021). Google Bert- Multiple Choice Question Generation on Ontology Base. *Journal of Cardiovascular Disease Research 12, 730-740.*

Hudon, A., Kiepura, B., Pelletier, M., & Phan, V. (2024). Using ChatGPT in Psychiatry to Design Script Concordance Tests in Undergraduate Medical Education: Mixed Methods Study. *JMIR Medical Education 10.* https://doi.org/10.2196/54067

Johnson, M., Ribeiro, A. P., Drew, T. M., & Pereira, P. N. R. (2023). Generative AI use in dental education: Efficient exam item writing. *Journal of Dental Education 87, 1865-1866.*

Kalpakchi, D., & Boye, J. (2021). BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset. *In Proceedings of the 14th International Conference on Natural Language Generation (pp. 387-403).* https://doi.org/10.18653/v1/2021.inlg-1.43

Khilnani, A. K. (2023). Potential of Large Language Model (ChatGPT) in Constructing Multiple Choice Questions. *GAIMS Journal of Medical Sciences 3, 1-3.* https://doi.org/10.5281/zenodo.7751267

Kic-Drgas, J., & Kılıçkaya, F. (2024). Using Artificial Intelligence (AI) to create language exam questions: A case study. *XLinguae 17, 20-33.* https://doi.org/10.18355/XL.2024.17.01.02

Kıyak, Y. S. (2023). A ChatGPT Prompt for Writing Case-Based Multiple-Choice Questions. *Revista Española de Educación Médica 4, 98-103.* https://doi.org/10.6018/edumed.587451

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology 80, 729-735.* https://doi.org/10.1007/s00228-024-03649-x

Klang, E., Portugez, S., Gross, R., Kassif Lerner, R., Brenner, A., Gilboa, M., Ortal, T., Ron, S., Robinzon, V., Meiri, H., & Segal, G. (2023). Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Medical Education 23.* https://doi.org/10.1186/s12909-023-04752-w

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice 41, 212-218.* https://doi.org/10.1207/s15430421tip4104_2

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education 30, 121-204.* https://doi.org/10.1007/s40593-019-00186-y

Laupichler, M. C., Rother, J. F., Grunwald Kadow, I. C., Ahmadi, S., & Raupach, T. (2024). Large Language Models in Medical Education: Comparing ChatGPT- to Human-generated exam questions. *Academic Medicine 99, 508-512.*

Laverghetta, A., & Licato, J. (2023). Generating Better Items for Cognitive Assessments Using Large Language Models. *In 18th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 414-428).*

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies 29, 11483-11515.* https://doi.org/10.1007/s10639-023-12249-8

Lee, J., Smith, D., Woodhead, S., & Lan, A. (2024). Math Multiple Choice Question Generation via Human-Large Language Model Collaboration. *In 17th International Conference on Educational Data Mining.* https://doi.org/10.48550/arXiv.2405.00864

Liang, Y., Wang, J., Zhu, H., Wang, L., Qian, W., & Lan, Y. (2023). Prompting Large Language Models with Chain-of-Thought for Few-Shot Knowledge Base Question Generation. *In 2023 Conference on Empirical Methods in Natural Language Processing (pp. 4329-4343).* https://doi.org/10.48550/arXiv.2310.08395

Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out, 74-81.*

Lopez, L. E., Cruz, D. K., Cruz, J. C. B., & Cheng, C. (2020). Transformer-based End-to-End Question Generation. *CoRR*. https://doi.org/10.48550/arXiv.2005.01107

Lu, O. H. T., Huang, A. Y. Q., Tsai, D. C. L., & Yang, S. J. H. (2021). Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students' Learning Performance. *Educational Technology & Society 24, 159-173.*

Maity, S., Deroy, A., & Sarkar, S. (2023). Harnessing the Power of Prompt-based Techniques for Generating School-Level Questions using Large Language Models. *In ACM International Conference Proceeding Series (pp. 30-39).* https://doi.org/10.48550/arXiv.2312.01032

Meißner, N., Speth, S., Kieslinger, J., & Becker, S. (2024). EvalQuiz – LLM-based Automated Generation of Self-Assessment Quizzes in Software Engineering Education: *Software Engineering im Unterricht der Hochschulen 2024.* https://doi.org/10.18420/seuh2024_04

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification 5, 181-204.* https://doi.org/10.1007/BF01897163

Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence 12, 1-32.* https://doi.org/10.1007/s13748-023-00295-9

Nasution, N. E. A. (2023). Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education 2.* https://doi.org/10.29333/agrenvedu/13071

Ngo, A., Gupta, S., Perrine, O., Reddy, R., Ershadi, S., & Remick, D. (2024). ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Academic Pathology 11.* https://doi.org/10.1016/j.acpath.2023.100099

Offerijns, J., Verberne, S., & Verhoef, T. (2020). Better Distractions: Transformer-based Distractor Generation and Multiple Choice Question Filtering. http://dx.doi.org/10.48550/arXiv.2010.09598

Olney, A. M. (2023). Generating Multiple Choice Questions from a Textbook: LLMs Match Human Performance on Most Metrics. *Grantee Submission.*

Onal, S., & Kulavuz-Onal, D. (2023). A Cross-Disciplinary Examination of the Instructional Uses of ChatGPT in Higher Education. *Journal of Educational Technology Systems 52, 301-324.* https://doi.org/10.1177/00472395231196532

Page, M., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology 134, 178-189.* https://doi.org/10.1136/bmj.n71

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311-318).* https://doi.org/10.3115/1073083.1073135

Rai, L., Deng, C., & Liu, F. (2023). Developing Massive Open Online Course Style Assessments using Generative AI Tools. *In IEEE 6th International Conference on Electronic Information and Communication Technology (pp. 1292-1294).*

Randolph, J. J. (2009). A Guide to Writing the Dissertation Literature Review. *Practical Assessment, Research, and Evaluation 14*. https://doi.org/10.7275/b0az-8t74

Rathi, S., Pophale, V., Kutwal, P., Mishra, S., & Nadkarni, A. (2024). Question and Assessment Generator: Deep Learning Approach for Customizable and Intelligent Assessment Creation. *In IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (pp. 618-624).*

Rivera-Rosas, C. N., Calleja-López, J. R. T., Ruibal-Tavarez, E., Villanueva-Neri, A., Flores-Felix, C. M., & Trujillo-López, S. (2024). Exploring the potential of

ChatGPT to create multiplechoice question exams. *Education Media 25.* https://doi.org/10.1016/j.edumed.2024.100930

Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models. *Expert Systems with Applications 208.* https://doi.org/10.1016/j.eswa.2022.118258

Sihite, M. R., Meisuri, M., & Sibarani, B. (2023). Examining the Validity and Reliability of ChatGPT 3.5-Generated Reading Comprehension Questions for Academic Texts. *Randwick International of Education and Linguistics Science Journal 4, 937-944.* https://doi.org/10.47175/rielsj.v4i4.835

Singh, M., Patvardhan, C., & Vasantha Lakshmi, C. (2023). Does ChatGPT spell the end of Automatic Question Generation research? *In IEEE International Conference on Computer Vision and Machine Intelligence (pp. 1-6).*

Soni, S., Kumar, P., & Saha, A., (2019). Automatic Question Generation: A Systematic Review. *In International Conference on Advances in Engineering Science Management & Technology.* https://dx.doi.org/10.2139/ssrn.3403926

Stadler, M., Horrer, A., & Fischer, M. R. (2024). Crafting medical MCQs with generative AI: A how-to guide on leveraging ChatGPT. *GMS Journal for Medical Education 41.* https://dx.doi.org/10.3205/zma001675

Suhartono, D., Majiid, M. R. N., & Fredyan, R. (2024). Towards automatic question generation using pre-trained model in academic field for Bahasa Indonesia. *Education and Information Technologies.* https://doi.org/10.1007/s10639-024-12717-9

Tran, A., Angelikas, K., Rama, E., Okechukwu, C., Smith IV, D. H., & MacNeil, S. (2023). Generating Multiple Choice Questions for Computing Courses Using Large Language Models. *In 2023 Proceedings - Frontiers in Education Conference.*

Tsai, D. C. L., Chang, W. J. W., & Yang, S. J. H. (2021). Short Answer Questions Generation by Fine-Tuning BERT and GPT-2. *In 29th International Conference on Computers in Education Conference (pp. 508-514).*

Ushio, A., Alva-Manchego, F., & Camacho-Collados, J. (2023). Generative Language Models for Paragraph-Level Question Generation. *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 670-688).* https://doi.org/10.48550/arXiv.2210.03992

Uto, M., Tomikawa, Y., & Suzuki, A. (2023). Difficulty-Controllable Neural Question Generation for Reading Comprehension using Item Response Theory. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 119-129).* https://doi.org/10.18653/v1/2023.bea-1.10

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *In 31st Conference on Neural Information Processing Systems.* https://doi.org/10.48550/arXiv.1706.03762

Vincentio, K., & Suhartono, D. (2022). Automatic Question Generation using RNN-based and Pre-trained Transformer-based Models in Low Resource Indonesian Language. *Informatica 46, 103-118.* https://doi.org/10.31449/inf.v46i7.4236

Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards Human-Like Educational Question Generation with Large Language Models. *Lecture Notes in Computer Science 13355, 153-166.* https://doi.org/10.1007/978-3-031-11644-5_13

Wolfe, J. H. (1976). Automatic question generation from text – an aid to independent study. *Proceedings of the SIGCSE-SIGCUE joint symposium on Computer science education 10, 104-112.* https://doi.org/10.1145/953026.803459

Wu, Y., Nouri, J., Megyesi, B., Henriksson, A., Duneld, M., & Li, X. (2023). Towards Data-Effective Educational Question Generation with Prompt-Based Learning. *Lecture Notes in Networks and Systems 711, 161-174.* https://doi.org/10.1007/978-3-031-37717-4_11

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 610-625).* https://doi.org/10.18653/v1/2023.bea-1.52

Appendix

Quality criteria		
Potential for distraction (quality of distraktors): 10, 11, 12, 20, 35, 44 **(6)**	Accuracy / specificity: 3, 5, 8, 11, 19, 23, 40, 43, 44, 51, 54 **(11)**	ROUGE-L score: 4, 7, 10, 11, 19, 22, 36, 37, 39, 48, 52, 55, 57, 59 **(14)**
Adherence: 15 **(1)**	Overall rating: 10, 43, 60 **(3)**	ROUGE-N score: 10 **(1)**
Appropriateness: 8, 13, 24, 25, 29, 39 **(6)**	Good Distractor Rate: 6 **(1)**	Difficulty level: 1, 3, 9, 13, 14, 22, 29, 30, 32, 33, 36, 39, 41, 47, 48, 49, 50, 56 **(18)**
Answerability: 11, 15, 22, 55, 56 **(5)**	Grammatical correctness: 3, 15, 16, 19, 21, 26, 36, 39, 43, 44, 55, 56, 58, 59 **(14)**	SMOG-Index: 60 **(1)**
Ease of use: 10, 35 **(2)**	Inter-rater reliability (Cohen´s Kappa, Fleiss´ Kappa*, Gwet´s AC1**): 1, 2*, 6, 11, 12* **, 14, 15, (34), 39 43*, 55* **(11)**	Fluency: 11, 28, 43, 47, 60 **(5)**
BERT score: 39, 55 **(2)**	Intra-rater reliability: 43 **(1)**	Question structure / organization / formatting: 14, 16 **(2)**
BLEU 1 score: 4, 7, 19, 22, 37, 48, 52, 57, 59 **(9)**	Clarity: 5, 12, 14, 25, 44, 47, 55 **(7)**	Toxicity (inappropriate content): 58 **(1)**
BLEU 2 score: 4, 7, 19, 22, 37, 48, 52, 57, 59 **(9)**	Coherence: 11, 22, 58, 31, 50, 60 **(6)**	Training Loss: 46, 54 **(2)**
BLEU 3 score: 4, 7, 19, 22, 37, 48, 52, 57, 59 **(9)**	Nonsense Distracor Rate: 6, 20 **(2)**	Discrimination index: 8, 9, 30, 32, 33, 41, 53 **(7)**

BLEU 4 score: 4, 7, 11, 19, 22, 36, 37, 48, 52, 55, 57, 59 **(12)**	Pedagigical value: 18, 23 **(2)**	Type/token ratios (TTR): 60 **(1)**
BLEU score (aggregated): 39, 46, 50 **(3)**	Plausibility: 26 **(1)**	Distinguishability oft the questions (manually created vs. LLM generated): 8, 9, 12, 13, 17, 24, 32, 38, 39, 43, 58, 60 **(12)** **(8)**
BLEU-Score (overall): 4, 7, 11, 19, 22, 36, 37, 39, 46, 48, 50, 52, 55, 57, 59 **(15)**	(Context)relevance oft he questions: 3, 4, 8, 10, 11, 15, 16, 18, 19, 21, 22, 24, 25, 28, 35, 36, 39, 44, 51, 56, 60 **(21)**	Validation Loss: 46, 54 **(2)**
ChrF score: 39 **(1)**	Cosine similarity: 48 **(1)**	Validity: 1, 11,17, 28, 33, 34, 35, 40, 41, 49 **(10)**
Suitability / usefulness: 8, 14, 15, 16, 29, 45, 51, 60 **(8)**	Cost efficiency: 18 **(1)**	Diversity: 2, 40, 58, 59 **(4)**
Novelty: 39 **(1)**	Student learning performance: 38 **(1)**	Truthfulness: 14, 18, 31, 40, 42, 53, 60 **(7)**
Reasoning ability: 1 **(1)**	METEOR score: 7, 11, 19, 22, 36, 37, 39, 55 **(8)**	Repetition score: 60 **(1)**
F1-Score: 19 **(1)**	More Over score: 55 **(1)**	Willingness to pay: 10 **(1)**
Felsch-Index: 60 **(1)**	Negative log-likelihood (NLL): 60 **(1)**	Time efficiency: 10, 18, 23, 28, 40 **(5)**
Formulation: 4, 18, 48 **(3)**	Reliability: 33, 41, 49 **(3)**	Time complexity: 23 **(1)**
Completeness: 2 **(1)**	ROUGE-2 score: 39 **(1)**	

Table 12: Assignment of Quality Criteria

* The numbers behind the quality criteria indicate in which of the studies (see Table 11) the respective quality criterion was used.

** The numbers in bold in brackets indicate the total number of studies in which the respective quality criterion was used.

Quality criterion (selection)	**Description**
BLEU score	(Automated) metric for evaluating the quality of machine-generated content. It measures the similarity between machine-generated texts and one or more reference texts by capturing how many word groups (n-grams) from an LLM-generated text are present in the reference text. The order in which these word groups appear is irrelevant. The BLEU score is always reported as a relative value, represented as a percentage or as a decimal number within the real-valued interval [0;1]. Accordingly, the scores range from 0 (no match) to 1 (perfect match). Higher BLEU scores indicate better quality translations that are closer to the reference texts in both linguistic and semantic terms (Papineni, 2002). However, strictly speaking, the BLEU score does not provide direct information about semantic similarity, even though literature suggests positive correlations between BLEU scores and semantic quality (Guan *et al.*, 2024). It should also be noted that multiple variations of the BLEU score exist, though only BLEU-1 to BLEU-4 were employed in the analyzed studies. While BLEU-1 is relatively superficial and only checks whether the correct words (or 1-grams) are used, BLEU-2 to BLEU-4 increase the rigor of evaluation by assessing whether the words appear in meaningful sequences. BLEU-4 is therefore considered the most stringent as it evaluates both word choice and structure. However, BLEU-4 is more sensitive to minor errors and synonyms.
Grammatical correctness	Refers to the linguistic accuracy of a task's formulations. This includes adherence to the grammatical rules of the respective

	language, such as the correct use of sentence structure, parts of speech, agreement (e.g., between subject and predicate), verb tenses, and punctuation.
Inter-rater reliability (IRR)	Indicates the degree of agreement in evaluation results between different raters. A high (or low) IRR signifies that the individual raters tend to strongly (or weakly) agree in their assessments regarding the evaluated quality criterion. This allows conclusions to be drawn about the reliability and consistency of the evaluation results.
Item Response Theory (IRT)	A mathematical model that represents the relationship between an individual's abilities and the probability of providing the correct answer to a given question.
Context relevance	Indicates how well an LLM-generated question is tailored to the specific requirements, conditions, and objectives of a given subject-specific context.
METEOR score	An (automated) metric for evaluating the quality of machine-generated content. It accounts for synonyms, word stems, and word order, enabling, in contrast to BLEU-1 to BLEU-4 and ROUGE-L metrics, insights into the semantic quality of LLM-generated texts (Denkowski & Lavie, 2014).
ROUGE-L score	An (automated) metric for evaluating the quality of machine-generated content. It measures the length of the longest common word sequence between LLM-generated texts and reference texts, represented as a relative value in percentage or as a decimal number within the real-valued interval [0;1]. Similar to the BLEU score, higher ROUGE-L values indicate stronger content and linguistic alignment (Lin, 2004). The key difference from the BLEU score is that ROUGE-L also accounts for the order of word sequences. However, it should be noted that the ROUGE-L score does not provide direct insights into the semantic quality of the text.
Difficulty level	Describes the relative difficulty of a task for the target audience being assessed. It serves as a measure of how well examinees can handle a particular task and indicates the level of challenge the

	task poses in terms of competence levels, subject knowledge, and cognitive demands.
Bloom´s taxonomy	A pedagogical model that systematically describes learning objectives and the cognitive abilities required to achieve them, classified along a multi-level scale (Krathwohl, 2002).
Discrimination index	Measures how well a given task is suited to differentiate between high-performing and low-performing examinees. Specifically, it indicates whether individuals who perform well (or poorly) on the overall exam also tend to perform well (or poorly) on the task in question (Tran *et al.*, 2023).

Table 13: Description of Selected Quality Criteria