Ripoll Y Schmitz and Sonnleitner
Large-scale Assessments in Education *(2025) 13:20*
https://doi.org/10.1186/s40536-025-00255-w

Large-scale Assessments
in Education

RESEARCH Open Access

Evaluating AI-generated vs. human-written reading comprehension passages: an expert SWOT analysis and comparative study for an educational large-scale assessment

Lisa Marie Ripoll Y Schmitz[1] and Philipp Sonnleitner[1*]

*Correspondence:
philipp.sonnleitner@uni.lu

[1] Luxembourg Centre for Educational Testing, University of Luxembourg, 11, Porte de Sciences, L-4366 Esch-Sur-Alzette, Luxembourg

Abstract

Background: The increasing capabilities of generative artificial intelligence (AI), exemplified by OpenAI's transformer-based language model GPT-4 (ChatGPT), have drawn attention to its application in educational contexts. This study evaluates the potential of such models in generating German reading comprehension texts for educational large-scale assessments, within the multilingual context of Luxembourg. Addressing the challenges faced by item developers in sourcing or manually developing numerous suitable texts, the study aims to determine if ChatGPT can assist text creation while maintaining high-quality standards.

Methods: The study employed a mixed-methods approach. In a qualitative focus group discussion, experts identified the strengths, weaknesses, opportunities and threats (SWOT) of using GPT-4 for text generation. These insights informed the construction of a Text Analysis Cognitive Model (TACM), which served as theoretical foundation. Narrative and informative reading comprehension texts were then generated using two distinct prompt engineering techniques, derived from original passages and TACM specifications. In a blinded online review, $N = 89$ participants evaluated human-written and AI-generated texts with regard to their readability, correctness, coherence, engagement and adequacy for reading assessment.

Results: All administered texts were of similarly high quality, with reviewers being unable to consistently identify authorship origins. Quantitative evaluations indicated that one-shot prompts are effective for creating high-quality informative texts, whereas human-written texts remain superior for narratives. Zero-shot prompts offer considerable flexibility and creativity, but still require human refinement.

Conclusion: These findings offer promising first insights into GPT-4's capacity to emulate human-written texts which can be used in the large-scale assessment context. The considerable potential of using generative AI-models as a flexible and efficacious assistant in the creation of reading comprehension texts is highlighted. Still, the necessity of human oversight is emphasized through an augmented intelligence-driven perspective. Given the jurisdictional framework of the European Union, an effective implementation of ChatGPT in the test development process remains hypothetical at this time but is likely to change.

Keywords: Generative artificial intelligence, Large language models, ChatGPT, Reading comprehension, Educational large-scale assessment, Text analysis cognitive model, Prompt engineering

Introduction

Large-scale educational assessment plays a pivotal role in the empirical monitoring of learning outcomes in students, allowing to objectively evaluate the performance, fairness, and long-term development of educational policies. With quickly developing advancements in educational technologies and ever-changing learning conditions, the demand for high-quality items in the context of standardized testing has increased continuously over recent decades (Tan et al., 2024). The drive for national and international benchmarks, such as the "Programme for International Student Assessment" (PISA) presents a significant challenge to the traditional labor-intensive and costly process of creating test items individually (Circi et al., 2023; Kosh et al., 2018).

As a potential strategy for streamlining the item development process efficiently, the technology-driven approach of Automatic (or Automated) Item Generation (AIG) has gained significant prominence across various areas and item types (Gierl & Haladyna, 2013; Haladyna & Rodriguez, 2013). AIG employs cognitive modelling as strong theoretical foundation, wherein subject matter experts (SMEs) specify the skills and abilities required to solve a test item correctly. These specifications are formulated on the basis of the SMEs' considerable expertise and familiarity with the examinees, their academic curriculum, and their learning environment. Subsequently, item features in the form of manipulable variables and values are defined within the notion of an item model, a schema, or a template. Computer-based algorithms are then used to instantiate these values, facilitating the prompt generation of numerous high-quality items with consistent characteristics. (Attali et al., 2022; Bezirhan & von Davier, 2023; Circi et al., 2023; Gierl & Haladyna, 2013; Gierl et al., 2020, 2021; von Davier, 2018). However, for the purpose of assessing reading comprehension, the conventional use of AIG is practically limited: A coherent text passage is required in addition to several (multiple-choice) questions aligning with its content, thus increasing the difficulty and complexity for defining the required item model's specifications (Attali et al., 2022; Sayin & Gierl, 2024; Xiao et al., 2023). Reading comprehension encompasses the extent to which examinees can appraise a reading passage's key messages and general sentiment, and their ability to draw inferences from the content (Attali et al., 2022; Sayin & Gierl, 2024), underlining the significance of the text to be read. Potential sources in the form of newspapers, magazines, textbook or children's book articles often require thorough adjustments and rephrasing when used for assessment purposes due to variations in topics, length, and inappropriate difficulty levels (Xiao et al., 2023). SMEs must rely solely on their expertise when meticulously identifying all relevant text features (words, phrases, sentences) and constraints that support the correct answer, within a self-written text passage or one sourced from the literature (Sayin & Gierl, 2024). The resulting reliance on highly skilled SMEs can entail substantial resource costs, as the cost-effectiveness of template-based AIG is contingent upon the assumption that the majority of generated items belong to the same content domain and that, consequently, they can be modeled based on the same cognitive model (Kosh et al., 2018).

In response to these issues, large language models (LLMs), such as OpenAI's family of GPT (Generative Pre-trained Transformer) models have attracted considerable interest in the context of AIG. These models only necessitate minimal fine-tuning without task-specific architectural adjustments to perform various natural language processing tasks, such as text generation (Bezirhan & von Davier, 2023). LLMs are designed to predict the next word sequentially based on the sentence's context by leveraging the initial input (prompt) and previously generated words. Human language inputs are encoded into rich, contextually embedded textual representations, which the system then decodes to produce coherent and context-appropriate responses (OpenAI, 2023; Tan et al., 2024). In their comprehensive review of AIG techniques leveraging LLMs in the educational field, Tan and colleagues (2024) concluded that LLMs offer a highly effective and flexible solution. They have proven helpful in generating large item banks, necessitating a few constraints regarding item type, language, the subject domain, or the data source used for further training. However, Tan and colleagues (2024) also highlighted the fact that many studies employing LLMs in AIG lacked a solid educational foundation. As a result, items were often generated without a thorough consideration of their measurement purposes and characteristics, which are crucial for educational assessment. Most studies did not consider the requisites of template-based AIG, which should begin with a clear definition of what, why, and how to measure, and thus failed to assess higher-level cognitive processes. In addition, SMEs have rarely been included in the AIG process, constituting a notable gap in post-generation item evaluation. However, the assurance of high item quality and their fulfillment of pedagogical demands is pivotal for their effective use in large-scale assessment. While most items were generated in English, the authors identified a total of 12 different languages for the generated items, suggesting the broad linguistic context in which LLMs can be used for AIG (Tan et al., 2024). In sum, while OpenAI's GPT-3 and GPT-3.5 models have been successfully leveraged for a range of educational applications, including automated (multiple-choice) question generation (e.g., Lee et al., 2023; Lin & Chen, 2024; Tomikawa & Uto, 2024; Wang et al., 2022), automated scoring (e.g., Jung et al., 2024; Latif & Zhai, 2024) and the provision of automated feedback to students (e.g., Pankiewicz & Baker, 2023), their utilization for the domain of reading comprehension texts remains less extensively explored (Bezirhan & von Davier, 2023; Xiao et al., 2023): The generation of items (in the sense of questions) remains the primary focus of previous research. The creation of reading passages is still in its infancy, despite being of equal or even greater relevance to the assessment of reading comprehension. One notable exception is the study by Bulut and Yildirim-Erbasli (2022), who explored the simultaneous generation of texts and related items for a reading comprehension assessment. Using OpenAI's GPT-2 and Google's T5 models, they achieved promising results; however, their findings also highlighted the clear need for human evaluation and revision of both texts and items.

The present paper aims to deepen our understanding of how large language models (LLMs) can support the development of assessment texts. Building on previous research, it examines the enhanced capabilities of more advanced models—such as OpenAI's ChatGPT/GPT-4—which benefit from larger training datasets and more sophisticated transformer architectures. Specifically, the study investigates their potential for generating reading comprehension stimuli suitable for use in a large-scale national assessment. We integrate this technology into the test development process by drawing from established template-based AIG standards through the development of a Text Analysis Cognitive Model (TACM, Sayin & Gierl, 2024). Emphasizing the integration of a solid theoretical foundation, SMEs will be involved throughout the entire development process to maintain high-quality standards. The resulting TACM serves as a "prompt template" considering constraints like word count or readability scores, along with input from previously developed assessment texts, so as to generate sentences and text passages efficiently (Tan et al., 2024).

Such template-based stimulus creation is especially important in the present context of the Luxembourgish national school monitoring program, the Épreuves Standardisées (ÉpStan). The distinctive multilingual orientation of the country's education system necessitates specific considerations regarding the quality and quantity of test development. For instance, the language of instruction in preschool is primarily Luxembourgish, or in some cases, French, but switches to German in grades one to six of primary school. In secondary school, French is gradually added as a language of instruction with its ratio to German depending on the school track (Men.lu, 2023). The legislation stipulates an annual assessment of student reading comprehension (amongst other competencies) in German and French in alignment with the national curriculum, being mandatory for all educational institutions within the country. Since instruction language may diverge from the first language of the individual student, it is imperative to consider the heterogeneous linguistic backgrounds of students in Luxembourg when designing items to test their competencies (Ugen et al., 2021): Students' language proficiency levels in German and French are not comparable to those of native speakers, hence, the creation and adaptation of reading comprehension texts is particularly labor-intensive and time-consuming, due to the complexity of the task.

The implementation of LLMs through TACMs ensures that the generated texts are aligned with the curriculum, thereby favoring the standardization of reading comprehension texts even across languages, ultimately having the potential to considerably reduce test development costs. In exploring this approach, this study contributes to filling a gap in the existing literature by leveraging GPT-4, as one of the most sophisticated and advanced LLMs at the time of the data collection (spring 2024), for generating reading comprehension texts. Given the Luxembourgish context and to investigate the generalizability of the model's potential to languages other than English, the reading comprehension texts will be developed in German. Therefore, first insights are offered, whether one of the latest LLM technologies could potentially be utilized as an assistive tool during the process of text development for the ÉpStan in Luxembourg and, more broadly, within the context of educational large-scale assessments. To provide a more precise framework for this study, the following section will elaborate on the theoretical concepts central to our investigation.

Advantages of GPT-4 relative to previous architectures

Transformer-based models overcome previous challenges through innovations in natural language processing (NLP), specifically the newly introduced attention mechanism in the transformer architecture (Vaswani et al., 2017). Here, the fine-tuning happens through reinforcement learning with human feedback, meaning that the control of the base model results from the application of precisely designed prompts (prompt engineering) to the instruct model in the post-training process (OpenAI, 2023; Ziegler et al., 2019). Given the implied adaptations and censorship, they produce more restricted yet contextually appropriate responses that are more suitable for educational purposes, where safety is paramount.

GPT-4, the latest iteration in the GPT model family, is rumored to have a training parameter count in trillions (e.g., Schreiner, 2023), marking a significant increase in the number of weights and leading to improved performance: While GPT-3.5's performance in a simulated bar exam resulted in the bottom 10% of participants, GPT-4 was capable to pass the test, scoring within the top 10% of test takers (OpenAI, 2023). Through better reliability, higher creativity, and the ability to process more nuanced instructions compared to previous models, OpenAI (2023) states that GPT-4 performs at a comparable level to that of humans, placing this model at the top of professional and academic benchmarks. Not only does GPT-4 achieve superior results compared to existing LLMs in English, but it also demonstrates a notable level of proficiency in other languages, as shown in a multiple-choice test that included translated professional and academic content of 57 different subjects. Another significant advance of GPT-4 is the substantial reduction of "hallucinations", referring to the generation of erroneous, inaccurate information (OpenAI, 2023).

As the volume of data utilized for training increases for every new model iteration, and with the incorporation of techniques such as retrieval augmented generation (RAG) and few-shot learning, LLMs capabilities are increasingly remarkable, matching or even exceeding human-level performance in various language processing tasks (Ackerman & Balyan, 2023; Brown et al., 2020; OpenAI, 2023; Tan et al., 2024; Xiao et al., 2023). These models demonstrate exceptional linguistic capabilities of producing grammatically correct sentences, making cultural references, and modelling complex contexts (Bezirhan & von Davier, 2023; Reynolds & McDonell, 2021). They are capable of learning a text's format and style from merely a few instructions or examples, thereby enabling the generation of novel, coherent and information-rich content without the necessity for explicit fine-tuning (Attali et al., 2022; Bezirhan & von Davier, 2023; OpenAI, 2023; Xiao et al., 2023).

For AIG in the educational sector, this represents an auspicious opportunity for innovative approaches, as it allows test developers to prototype new item types and iteratively refine them without having to undertake a lengthy model development process (Attali et al., 2022; Xiao et al., 2023). Given that GPT-4 is widely regarded as one of the most sophisticated and powerful LLMs (Bezirhan & von Davier, 2023), it is evident that there have been considerable improvements in terms of reliability and accuracy of the outputs generated. Still, it is impossible to rule out errors in the logical reasoning process. The challenge of ensuring the correctness and trustworthiness persists, especially in high-stakes contexts, such as educational assessment. Consequently, human involvement continues to play a pivotal role in this regard.

Effectiveness of zero-shot vs. one-shot prompting

OpenAI's ChatGPT is only capable of generating personalized reading passages based on carefully crafted prompts, rendering them highly important. As the autoregressive model draws upon GPT-4 architectures, which are complemented by human instructions to enhance its performance and response accuracy significantly, the design of the prompts exerts a direct influence on the overall quality and relevance of the outputs generated by the model (Bezirhan & von Davier, 2023). Here, the term prompt engineering refers to the construction of commands or instructions that effectively communicate tasks to LLMs (Tan et al., 2024), with varying degrees of control over the output: In *zero-shot prompts*, or zero-shot generation, only a natural language description is provided without the explicit demonstration of an example. This encourages creativity and flexibility in the model's responses, as it solely relies on its pre-acquired knowledge. In the case of *one-shot prompts*, or one-shot generation, an example is provided in addition to the manual instruction. This facilitates the model's learning of specific linguistic styles and structural features, as a generated text passage will be similar to the given reference (Bezirhan & von Davier, 2023).

A review of existing research indicates that the one-shot method may offer greater efficiency than the zero-shot method in specific contexts, as LLMs are generally estimated to be few-shot learners (i.e. requiring several examples) (Brown et al., 2020). In their study on automated reading passage generation with OpenAI's GPT-3, Bezirhan and von Davier (2023) demonstrated that integrating specific information about the target group's grade was the most effective method in matching the constrained topic and text difficulty score to the original text. These results for one-shot learning align with prior research, indicating better performance with clear instructions and multiple examples (Brown et al., 2020; Reynolds & McDonell, 2021). Conversely, the zero-shot prompts exhibited unsatisfactory results, as GPT-3 demonstrated superior performance when clear instructions are provided during the text generation process. An interpretation of the consistently superior performance of GPT-3 with the provision of examples is that these examples serve simply to instruct the model on the task to be solved, encouraging it to adhere to the structure of the prompt in its response. Reynolds and McDonnell (2021) illustrated this assumption through a conducted French-to-English translation task, demonstrating that zero-shot prompts could achieve comparable and superior results to those obtained through the few-shot format. Given the extensive array of functions that do not require learning at runtime, zero-shot prompting could facilitate greater adaptability and originality. For instance, Xiao and colleagues (2023) found in their study that English reading passages generated by ChatGPT were consistently rated higher in quality for both zero-shot and one-shot settings than those written by humans, with zero-shot passages still scoring slightly higher. It can be inferred that alternative prompts without examples or tailored instructions, thus corresponding to zero-shot or baseline performance without meta-learning, are significantly underestimated (Reynolds & McDonell, 2021). Given the context of the present study, it seems crucial to consider both prompting strategies since superiority remains disputed to date.

Augmented intelligence through cognitive model-based prompt engineering

Given the often-elevated expectations coming along with such innovative technologies, the present study aims to realistically appraise GPT-4's capabilities within the overarching framework of (Hybrid) "Augmented Intelligence" (Gierl et al., 2020; Zheng et al., 2017). This field of artificial intelligence is concerned with the extension of human cognitive abilities, by combining modern AI's strengths of normalization, repeatability, and logicality with the human mind's creativity, complexity and flexibility (Gierl et al., 2020; Sayin & Gierl, 2024; Zheng et al., 2017). The perspective of human-centered AI integrates AI under human control, thereby fostering productivity with higher reliability, safety, and trust and mitigating potential risks for high-stakes contexts (Yang et al., 2023).

The TACM, as first introduced by Sayin and Gierl (2024), exemplifies this approach of strategic collaboration for addressing more complex issues, such as reading assessment in educational monitoring. It is fundamentally based on the three-stage process of AIG as outlined by Gierl and Lai (2013): Firstly, SMEs identify the relevant content; secondly, this content is integrated into the cognitive model (CM) and transferred into an item template; in the final step, the item generation process is initiated. In the context of an augmented intelligence-driven approach to test development (Sayin & Gierl, 2024; Zheng et al., 2017), the TACM additionally leverages the generative abilities of GPT. Precisely, the item model instructions created by a SME serve as instructions (prompts) for GPT to produce text content for reading comprehension items. Sayin and Gierl (2024) successfully implemented the TACM to generate 12.500 items with GPT-3.5 for a reading comprehension task that required test takers to identify the irrelevant sentence within a total of five sentences.

Instead of generating distinct sentences, the present study utilizes the TACM as a theoretically grounded prompt template to structure key cognitive and linguistic constraints during text generation. This not only addresses the difficulties in identifying the features and constraints needed for text generation in template-based AIG but also enhances the robustness and applicability of the TACM for diverse reading comprehension contexts, and even languages. As the TACM draws upon cognitive models that human item developers have deliberately created, it demonstrates how the integration of AI technologies and human expertise can provide an efficacious and versatile solution for facilitating specialized content production while at the same time ensuring validity (Sayin & Gierl, 2024).

Discernibility of human-written vs. AI-generated texts

Finally, the question arises as to whether the generated texts are of such an elevated quality that they can no longer be distinguished from human-written texts. Especially in the context of educational assessment, acceptance of generated texts by students, teachers, as well as educational policy makers is crucial. Xiao and colleagues (2023) demonstrated that evaluators often perceived ChatGPT-generated passages as even more human-like than actual human-written passages. This effect was more pronounced in the one-shot setting, potentially due to the model effectively imitating the style and structure of the provided reference passage. Another study investigated the ability of experienced reviewers and linguists to discriminate between ChatGPT-generated and human-written scientific abstracts (Casal & Kessler, 2023). The findings indicated that participants demonstrated limited capacity to accurately identify all four of the presented abstracts, with the majority only able to identify one or two abstracts correctly. Interestingly, the reviewers were more successful at identifying the human-written texts than the AI-generated ones. This may suggest a heightened discernibility towards human text, although it is not yet sufficient to ensure consistent and accurate identification (Casal & Kessler, 2023). In conclusion, it can be stated that ChatGPT can reliably mimic the style and structure of human writing.

Methods

The present paper employs both qualitative and quantitative approaches. In Study 1, a qualitative focus group discussion was conducted with an expert panel to identify the strengths, weaknesses, opportunities, and threats (challenges) of the proposed approach to generate reading comprehension texts with GPT-4. These outcomes served as methodological foundation for the TACM construction, used as a basis for the subsequent text generation process. Study 2 involved the generation of the reading comprehension texts based on prompt engineering per the content specifications and constraints defined within the TACM. Native German speakers, teachers, and test developers partook in an online survey to obtain a quantitative quality evaluation of the original and the generated texts. Through a blinded review, the dimensions of readability, correctness, coherence, engagement, and adequacy of the administered texts were rated to explore their potential for national school monitoring. Finally, participants were requested to provide a rating whether, based on their subjective judgement and antecedent quality assessment, they estimated the respective texts to be humanly authored or generated by artificial intelligence.

Study 1: Expert panel focus group discussion

Aim and research question

The rationale for employing a focus group discussion as the methodology for the expert interview was supported by its practical and dynamic way of simultaneously determining the perspectives of multiple respondents. Group discussions facilitate the multiplication of ideas, as listening to the answers of the other participants creates interaction and mutual stimulation, thus contributing to the conversation's depth and diversity (Bortz & Döring, 2006). The primary objective of this qualitative study was to gain a preliminary understanding of the attitudes and opinions, that SMEs engaged in test development might hold regarding the utilization of generative AI in language assessment. The experts working at the Luxembourg Centre for Educational Testing (LUCET), who are responsible for developing items used in the national school monitoring (ÉpStan), was of particular interest for the research objective. Given the specificity of the multilingual context in Luxembourg coupled with the novel field of utilizing LLMs for text generation, the focus group discussion was particularly suitable for exploring relevant issues prior to a more quantitative approach (Howitt & Cramer, 2017). The findings obtained in this initial phase were considered for methodological derivations regarding the subsequent TACM prompt design.

Hence, this step was implemented in order to ensure the generation of high-quality and appropriate textual outputs. The following research question can be derived from this objective:

Research Question 1 How do SMEs appraise the use of generative artificial intelligence (GPT-4) as a support tool in the context of language assessment?

Participants and procedure

To answer the first research question, it was essential for the participants to actively engage in the group discussion by collaboratively exploring and collectively fostering the generation of ideas about specified subtopics. The researcher solely moderated, ensuring that the pre-planned questions from the interview guide were covered and the time restrictions were met. Without actively participating, conversational drifts were reoriented towards the respective questions. The discussion lasted approximately one hour and was held in English.

The panel was comprised of a homogeneous group of $N=6$ test development experts, recruited from within LUCET. Five specialists in educational measurement at different grades, with expertise in reading and listening comprehension tasks in French, German or Luxembourgish languages, and one expert responsible for coordinating the related test development participated in the discussion. The panelists were well established in their respective fields, with extensive working experience in test development and item writing.

Measures

To conduct the expert panel focus group discussion, a semi-structured interview including seven questions (see Table 1) was designed (Breen, 2006). Additionally, the interview guide was conceptualized with consideration of the strengths, weaknesses, opportunities, and threats (SWOT) analysis, derived from the field of strategic management and marketing (Leigh, 2009). Including these four dimensions as guidelines for the discussion aimed at opening up a debate and facilitating comprehensive responses to the first research question, all while encouraging both favorable as well as more critical commentary.

Table 1 Semi-structured interview guide for focus group discussion

Component of SWOT analysis	Question
Strengths	1. Where do you see the strengths of incorporating AI in generating text passages for reading comprehension?
	2. Are there potential differences for the respective language?
Weaknesses	3. What could be potential weaknesses of this idea?
	4. What aspects should be kept in mind in the process of generating the texts?
Opportunities	5. Supposing that AI is capable of consistently producing high-quality texts, which opportunities do you foresee?
Threats	6. Where do you see potential threats?
	7. What are fears or criticisms which might arise in the use of AI for generating texts?

Results

The analysis of the focus group discussion was conducted through an inductive categorization of the data material (cf. Mayring, 2003), whereby recurring themes and patterns were identified. The procedure commenced with an initial listening to the audio recording, which was then literally transcribed into text. At this preliminary stage, the researcher noted first impressions about emerging themes, group dynamics and recurring patterns. Repeatedly listening to the audio recording and reading the script allowed for an increased familiarization with the content and established the foundation for an iterative and reflective approach to the subsequent thematic analysis (Howitt & Cramer, 2017). By writing down the answers in parallel, it was possible to avoid repetition and to quantify the prevailing themes of most significant relevance and incidence within the group (Howitt & Cramer, 2017). Particular attention was paid to the degree of consensus and alignment between the item developer's statements. The analysis was conducted in a systematic manner, with each question of the semi-structured interview guide and its respective themes being examined in turn. The aspects of "weaknesses" and "threats" were analyzed together, as these questions are semantically close, and the answers were pertinent for both categories. The use of a few illustrative examples adds to the descriptions and serves to emphasize the relevance of the identified categories in the context of the first research question: *How do SMEs appraise the use of generative artificial intelligence (GPT-4) as a support tool in the context of language assessment?*

Strengths of incorporating AI in generating text passages AI can provide a valuable first draft, thereby saving time in the text-creation process and offering a starting point that can then be built upon. In the case of developing complex and lengthy texts used in the evaluation of reading comprehension, AI can be a support by enabling more efficient content adaptations, as well as reducing syntactic and semantic complexity. With its extensive information database, AI can generate synonyms, simplify sentences, and introduce new ideas, whilst maintaining the original author's personal writing style. The desired textual outputs can be adapted according to specific topics and language proficiency levels, making them modifiable and suitable for varying student requirements. As AI-generated texts are not considered to be the intellectual property of one member in the test development group, they can be more easily subjected to open criticism and improvement during teacher meetings.

> *In German Grade 3 and 5 the original texts are mostly too long, too complex and also the texts are not written for the purpose of reading comprehension, but in order to change that it takes normally a few days to do it appropriately [...] so this might definitely be a help to save some work.*
> *We often have good ideas for potential topics, but we cannot find an appropriate text easily, so we would like to feed AI certain information about our idea, see what it comes up with and then rework some of the phrases.*

Potential weaknesses and threats The legal framework surrounding the use of AI-generated content raises significant concerns regarding intellectual property, authorship, plagiarism, and the legality of using AI in content creation. Furthermore, the unique nature of human creativity is a critical aspect mentioned. AI is incapable of generating truly creative content, as it merely replicates based on its input, which may result in loss of quality and the language's richness. It is, therefore, challenging to translate the essence of the human mind into corresponding constraints that AI can effectively replicate. This specificity required for text generation can also be time-consuming, making manual writing still more efficient. Lastly, as AI may produce inaccurate information, human verification and interpretation of the content is paramount.

> *That's the biggest threat, you're taking the intellectual property of somebody, because the internet is full of it, and publish it as your own.*
> *In the end, language is way more than that, it's what humans produce actively in their mind.*

Future opportunities of using AI to generate text passages Similar to improvements seen in online translation tools, AI is expected to match or exceed current human capabilities in text generation. From a psychometric perspective, as AI might be able to generate a large number of text versions with consistent constraints and levels of difficulty, these could serve for parallel testing. While text generation might be more straightforward, another opportunity lies in the potential standardization of item (in the sense of questions) development through AI. Lastly, AI challenges item developers to improve their skills, encouraging continuous professional development and the expansion of personal horizons.

> *I think, in the end, AI can to everything we can do in a few years [...], for now it has more information than each of us [item developers] can ever access or process as an individual.*
> *We have a tendency of always using the same vocabulary, the same syntax, but that can give us some ideas, broaden our horizons and maybe take routes that we would not have taken on our own.*

Potential language differences The scarcity of linguistic training data for Luxembourgish may pose a challenge for the capacities of AI, as this raises concerns about potential vocabulary biases and inadequacies of the AI-model. Additionally, cultural bias in the context of language technology is an important consideration, given that AI may face limits in understanding cultural connotations and nuances, as well as specific meanings attached to language. This may lead to difficulties in translating pragmatic aspects across languages.

> *In linguistics, pragmatics are very difficult to translate, sometimes almost impossible, as they are so culturally-bond to language.*

Aspects to keep in mind during text generation Verifying the accuracy and coherence of AI-generated texts requires a thorough proofreading process. It is the responsibility of all item developers involved in the creation of these texts to ensure that they are accurate, coherent with the intended message, and free of biases or other forms of manipulation that might be introduced during the text generation process. Item developers must assume this responsibility for the ultimate content, recognizing that AI has limitations and resisting the temptation to accept well-worded AI outputs without proper scrutiny. At present, the majority of AI-algorithms are primarily based in the United States and subsequently translated into other languages, making them more accurately calibrated for the English language.

> *You still have the responsibility to read it afterwards and see if the main message is still there, you really have to pay attention if the information is correct […] as quality always needs time and investment.*
> *I think it's tricky, because if the wording looks good, well, it's tempting to trust it […] and just overtake it, and to not read properly anymore because you give it a certain credibility.*

Discussion

To summarize and answer the first research question, the general sentiment of the focus group discussion paints a somewhat optimistic image of considering LLMs as an assistive tool in text development. The statements are consistent with the literature, underlining that it is difficult to find appropriate references for the development of reading comprehension texts and that this task area proves to be particularly labor-intensive (e.g., Sayin & Gierl, 2024; Xiao et al., 2023). Also, advantages applicable to the TACM were cited in that, on the one hand, efficient modifications or content adjustments can be made according to grade requirements and, on the other hand, parallel test versions with similar characteristics can be created. However, human creativity and control remain crucial to ensure the quality and appropriateness of the output content. The potential may also depend on the type of text and topic: the experts were not necessarily in agreement as to whether the text quality would be higher in informative (factual) texts due to doubts about the fictional creativity of the algorithm, or for narrative (literary) texts, as facts in informative texts are more likely to be incorrect. It was, therefore, decided to have both types of reading comprehension texts in the subsequent generation phase for comparison. Considering the grade level to be selected, the item developers discussed whether the text generation might be more difficult for younger children, where texts are written for a highly distinctive context, making it easier for secondary school children, as there would be fewer constraints. In contrast, it was argued that AI-generated texts may not provide enough depth for older students, as in grade nine it is important to go beyond the text, make inferences, or "read between the lines". Thus, the fifth grade as "medium" ÉpStan grade was chosen as the target group for the textual requirements. As one of the main conclusions, all interviewees insisted on the importance of SMEs' involvement in the pre-, during- and post-generation process of items and their responsibility for the final content, similar to the suggestions of Tan and colleagues (2024) and the augmented intelligence approach in general.

Study 2: Prompt engineering, text generation, and quality evaluation questionnaire

Aim and research question

The reading comprehension texts were developed to assess two key reading abilities, similar to Bezirhan and von Davier (2023): (1) literary experience, and (2) acquiring and using information. The former is mostly assessed through reading fictional text, while the latter is associated with understanding informative articles or instruction manuals. The methodology employed for text generation relied upon priming an LLM with a specific context, such as the constraints of the TACM, and then allowing the model to develop an independent text (Bezirhan & von Davier; Reynolds & McDonell, 2021). The generation process was inspired by the specific requirements of the narrative or informative text template, and either provided with a topic and additional example text (one-shot prompting) or not (zero-shot prompting). Based on existing literature, both approaches have proven to be effective with GPT-3 or GPT-3.5, contingent on the purpose. The manual design of natural language input prompts, based on professionally defined text model requirements, was intended to generate high-quality reading passages without additional control methods (Xiao et al., 2023). The effectiveness of the two prompt design approaches was quantified through the assessment results of the different quality dimensions. Hence, the design of the questionnaire employed in the present study was guided by the necessity to reliably and consistently assess the quality of the texts generated with GPT-4 through intrinsic human evaluation. A further aim was to ascertain the suitability of the texts for the reading-levels of fifth-graders in Luxembourg and gain a first impression about their potential for large-scale educational assessment. Given the absence of prior studies on the quality of German reading comprehension texts produced by the GPT-4 model, the following overarching research question is formulated:

Research Question 2 How do texts generated by different methods (human-written, zero-shot, one-shot) compare in terms of quality dimensions and their suitability for national education assessment?

Furthermore, an exploratory inquiry was included to elicit subjective judgements from respondents regarding whether the texts were created by an expert test developer or by artificial intelligence. Previous research indicates that reviewers were unable to accurately discriminate texts authored by ChatGPT from those composed by humans (e.g., Casal & Kessler, 2023; Xiao et al., 2023). This investigation aims to appraise the potential of GPT-4 as a support tool in text development and to discern qualitative distinctions between the two text development approaches. Hence, the research question is articulated as follows:

Research Question 3 Can (experienced) reviewers discern whether the texts presented were humanly authored or generated by artificial intelligence?

Participants and Procedure

First, the constraints for the TACM framework were established through collaboration with an SME from the LUCET. They were derived from both her expertise and the a priori written narrative and informative sample texts for the Luxembourgish fifth grade. The derived requisites in the TACM were then utilized as prompts to instruct GPT-4 on the text generation (see Appendix A). The creation of natural language requests was initiated with the objective to develop bespoke CustomGPT models within the interface of ChatGPT (based on GPT-4). These were tailored to narrative and informative text generation and categorized as either zero-shot or one-shot prompting, resulting in four distinct CustomGPT models (see Table 2). One significant advantage of a CustomGPT

Table 2 Overview of CustomGPT models and texts selected

	Human-Written	**Zero-Shot**	**One-Shot**
Narrative Text	NT-HW "Little monkey"	**NT-ZS** "Treasure hunt"	**NT-OS** "First day of school"
Informative Text	IT-HW "Wales"	**IT-ZS** "Fireflies"	**IT-OS** "Earthworms"

The four created CustomGPT models are highlighted in bold

model is that, if necessary, individual specifications may be updated at any time (Keyser, 2023) without having to start over.

Rather, the generation of text samples can be achieved with remarkable efficiency, requiring only a single mouse click within a user-friendly interface. Concurrently, other item developers may also access the corresponding CustomGPT model via a private link, thereby facilitating efficient and cooperative collaboration.

The natural phrasing of the prompts was based on OpenAI's official prompt engineering guide (OpenAI, n.d.) and additionally informed by previous studies in that area (Bezirhan & von Davier, 2023; Sayin & Gierl, 2024; Xiao et al., 2023). Assuming a higher output quality, due to GPT-4 being a multilingual model that is able to interpret and understand prompts written in 26 different languages (OpenAI, 2023), the texts were directly prompted in German. Several reading texts were then generated for each of the prompt designs. The final selection of text passages (see Table 2 and Appendix B) was conducted by an SME, who verified that they met the predefined criteria of the TACM and corresponded to the original text to a satisfactory extent. In contrast to the methodology in other research studies, the SME did not undertake any revision or editing of the text passages prior to the quality evaluation. This approach permitted the unaltered evaluation of potential grammatical or factual errors within the generated texts in their original form, thereby preventing any distortion of the quality assessments in the subsequent online questionnaire.

The online survey was conducted using the open-source survey tool LimeSurvey (LimeSurvey GmbH, n.d.) and spanned a period of three weeks. The participants were recruited through snowball sampling via various channels. In the survey, a total of six reading comprehension texts were administered, two of them being written by a human test expert and four texts generated by GPT-4, as based on the SME's selection. They were presented in a randomized order to prevent the occurrence of succession effects, and it was not indicated whether the texts were generated automatically or written by an expert. The order of items was alternated to encourage attentive processing, and to minimize the influence of response bias. The total completion time for the questionnaire took approximately fifteen minutes. All statistical analyses were conducted in RStudio (RStudio Team, 2020).

A total of $N=161$ review participants initially enrolled in the survey. Of these, 72 participants (44.2%) did not complete the study, resulting in a final sample size of $N=89$ participants (55.28% completion rate). An analysis of the drop-out data indicated no systematic differences between those who completed the study and those who did not, in terms of key demographic variables and the point of drop-out. This suggests that no systematic bias due to drop-out is expected to influence the study results and that the reason for attrition might be primarily due to loss of interest. Between the remaining $N = 89$ reviewers, 56 (62.92%) identified as female, 32 as male (35.96%), and one as non-binary (1.12%). They were on average 35.92 years old ($SD = 14.78$, range 18–77 years). The majority of participants reported having the German (70.79%) or Luxembourgish (20.22%) nationality, went to school in Germany (70.79%) or Luxembourg (17.98%) and grew up speaking German (73.03%) or Luxembourgish (14.61%) at home. It is therefore presumed that the sample was capable of adequately evaluating German reading comprehension texts in the context of Luxembourgish primary education. Additionally, most participants held either a Bachelor's degree (30.34%) or had completed a Master's degree, equivalent diploma or state examination (34.83%). In terms of profession, the majority indicated to be currently occupied as university student (42.70%) or employee (35.96%). 34 reviewers (38.20%) have had experience with teaching or educational test development, therefore constituting the expert subgroup relevant for the subsequent analysis of response differences.

Measures

The developed TACM is presented in Fig. 1 and comprises three levels of information (Sayin & Gierl, 2024). In the first-order level, the primary objective is the identification of both the underlying problem and the associated scenarios that are to be assessed in the reading comprehension task. As mentioned, the present study ascertains the examinee's capacity to discern the integrity of meaning in a text passage (cf. Sayin & Gierl, 2024) through either a scenario of literary experience or the utilization and acquisition of information. At the second-order level of the cognitive model, the source of information is specified; in this case, a German fictional (narrative) or factual (informative) text passage, both authored by an SME. The third-order level of analysis focuses on the main features of the text, which include semantic, organizational, and textual information. The semantic feature pertains to the general idea of the text that contributes to the comprehension of its' meaning and content. The organizational feature relates to the structure and expression of the content, which determines the arrangement of information within the text and the linguistic design used to convey it. For the narrative text passage, the SME specified clearly defined structural and linguistic constraints—such as paragraphing, direct speech, and turn-taking—which were applied as prompt conditions, whereas for the informative text passage it was essential to include titles for each paragraph describing animal characteristics. Lastly, the text feature concerns the desired number of words as well as the readability, which is adapted to the target group in terms of its lexical complexity and syntactic structure. Here, the SME specifically defined the subordinate conjunctions and sentence connectors to be used in a reading comprehension text for a Luxembourgish fifth grade. As most pre-existing readability indexes (e.g., Flesch-Kincaid, Wiener Sachtextformel) were developed primarily for English or German native speakers, they were deemed to be unsuitable for cross-language transfer to Luxembourgish students.

Consequently, these indexes were not integrated into the readability parameters of the TACM to prevent any distortion of results and ensure that the generated text remained within an appropriate difficulty range. Instead, it was postulated that the designated

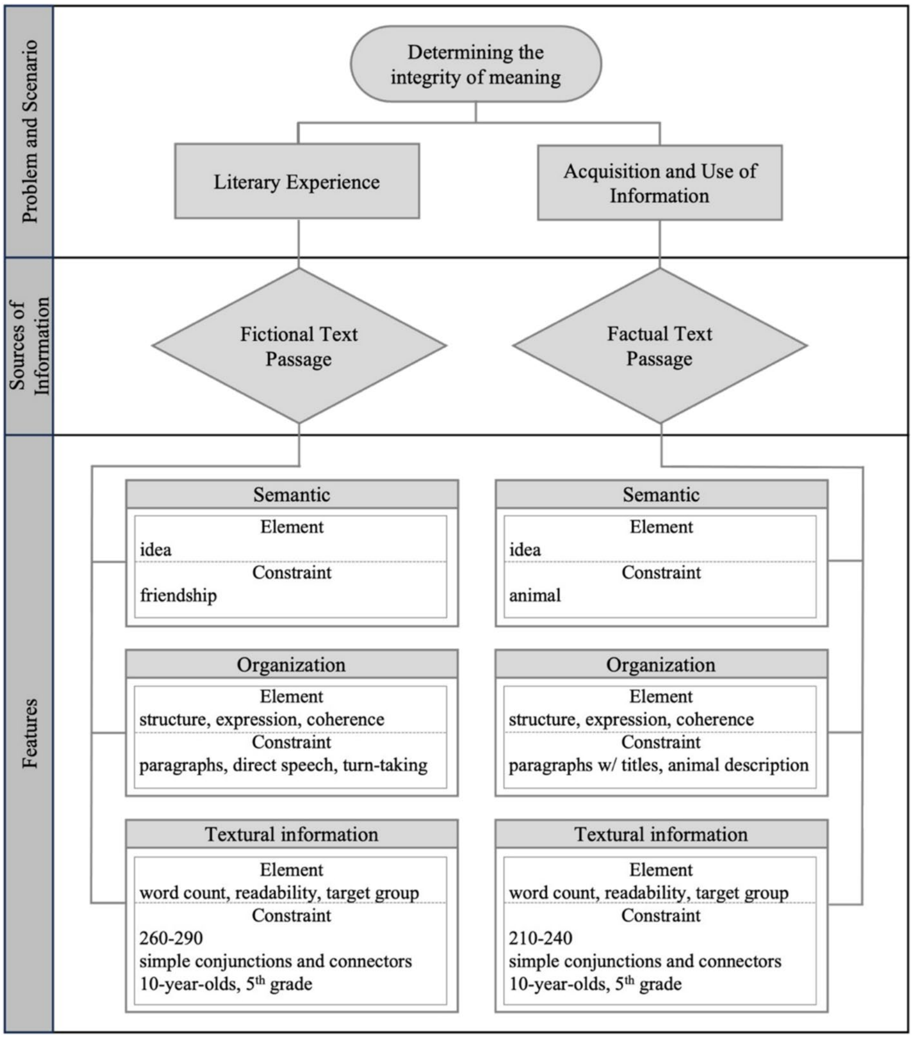

Fig. 1 Text analysis cognitive model (cf. Sayin & Gierl, 2024*)*. The constraints are based on the requirements defined by SME for Luxembourgish 5th grade

readability prompts, which include a word limit, the sample text (for one-shot prompting), and the target group's age and grade level, would sufficiently regulate the readability within the output texts. Each of the aforementioned features is divided into two nested components: elements and constraints. The original texts exhibit identical semantic, organizational, and textual characteristics; however, the respective constraints differ. The latter are necessary at all three levels, in order to ensure that the generated texts remain cognitively aligned, consistent, and comparable in difficulty, while still allowing diversity within the defined parameters (Sayin & Gierl, 2024).

The majority of text generation tasks are still subjected to quality assessment through intrinsic human evaluation, which continues to represent the gold standard for revision (Celikyilmaz et al., 2020). The most frequently employed evaluation techniques are Likert and sliding scales (Celikyilmaz et al., 2020). The evaluation of generated text in the academic literature encompasses both qualitative and quantitative methodologies. For instance, in a study conducted by Sayin and Gierl (2024), three SMEs assessed the quality of generated items using categories such as "Acceptable", "Minor Revision", "Major Revision", or "Reject" prior to a field test. In a similar vein, Lee and colleagues (2023) evaluated the clarity, utility, and comprehensiveness of protocols via eight English teachers from different teaching backgrounds, using a 44-item Likert scale and open-ended qualitative insights. Säuberli and Clematide (2024) investigated the answerability and guessability of multiple-choice German reading comprehension questions, including but not limited to, a rating scale from 1 "unusable" to 5 "perfect", involving six university students or recent graduates and native German speakers. Xiao and colleagues (2023) conducted a multidimensional quality assessment of generated English reading material, comprising the dimensions "readability" (easy and fluent to read), "correctness" (accurately reflects facts, common sense, and is logical), "coherence" (consistent with topics and storylines), "engagement" (interesting and engaging) and "overall quality", as well as a pairwise comparison of generated versus human-written texts. Here, nine college students and 364 native English speakers conducted the evaluations. The main author of this paper was generous in providing more in-depth explanations of their items, which formed the basis for the development of the present questionnaire. Bezirhan and von Davier (2023) assessed the coherence and appropriateness of GPT-3.5 generated texts for fourth-graders through an online survey, involving 50 expert reviewers. To do so, they employed the dimensions "adequate", "coherent", "engaging", "main topic" and "not distracting". Similarly, Attali and colleagues (2022) evaluated an interactive reading task created by GPT-3, with 18 SMEs undertaking a content (cohesion, clarity and logical consistency) and fairness (cultural specificity, technical or field specific jargon and sensitivity) review. In light of these methodological approaches, the target group for the evaluative scope of the present study not only encompassed SMEs, such as test developers and teachers. Rather, it was extended to participants having an advanced understanding of the German language, thereby facilitating a more numerous and comprehensive assessment of the generated texts with regards to their linguistic and cultural nuances.

The present text evaluation encompassed fifteen statements on five dimensions, derived from the aforementioned studies: "readability", "correctness", "coherence", "engagement", and "appropriateness". Participants were to indicate their accordance to each statement using a five-point Likert scale, ranging from 1 "strongly disagree", 2 "disagree", 3 "undecided", 4 "agree" to 5 "strongly agree". Moreover, respondents were given the option of providing additional comments in an open text field and were finally asked to provide an intuitive judgement as to whether the respective text passage appeared to have been authored by a human expert or generated by AI. The full item catalogue can be found in Appendix C.

Results

Overall results Regarding the results of the overall quality assessment, the six texts were generally well received and of satisfactory quality, with all mean scores being greater than four on the five-point Likert scale. The informative one-shot text ("Earthworms") exhibited the most favorable overall performance, achieving the highest mean score ($M = 4.45$, $SD = 0.51$). The mean value for the informative human-written text ("Wales") was 4.28 ($SD = 0.60$), shortly followed by the informative zero-shot text ("Fireflies"; $M = 4.21$, $SD = 0.66$). With regard to the narrative texts, the one authored by a human expert ("Little monkey") performed best ($M = 4.41$, $SD = 0.52$). The narrative zero-shot text ("Treasure hunt") demonstrated slightly inferior performance ($M = 4.20$, $SD = 0.66$), with the narrative one-shot text ("First day of school") receiving the lowest score overall ($M = 4.11$, $SD = 0.60$). Results for the specific quality dimensions of each text type are presented below.

Narrative texts A multivariate analysis of variance (MANOVA) was conducted for the category of narrative texts to examine the effect of text type (human-written, zero-shot, one-shot) on five dependent variables: readability, correctness, coherence, engagement, and adequacy. Using Pillai's Trace, there was a statistically significant main effect of text type on the quality ratings, $V = 0.169$, $F(10, 522) = 4.82$, $p < 0.001$. The post-hoc Tukey HSD tests, based on the significant univariate ANOVA analyses, revealed significant differences between the text types in all quality dimensions, except for correctness (see Fig. 2). The mean difference in readability between human-written and one-shot was − 0.31, 95% CI [− 0.54, − 0.08], $p = 0.005$. Between human-written and zero-shot, the mean difference in readability was − 0.41, 95% CI [− 0.64, − 0.18], $p < 0.001$. Hence, the human-written text had significantly higher readability compared to both the one-shot and zero-shot text. Concerning the elicited engagement, the human-written text showed significantly better results compared to one-shot (− 0.54, 95% CI [− 0.86, − 0.23], $p < 0.001$), while the zero-shot text had significantly higher engagement compared to one-shot (0.52, 95% CI [0.21, 0.84], $p < 0.001$. Furthermore, human-written text showed significantly higher coherence compared to the one-shot text (−0.31, 95% CI [− 0.55, − 0.06], $p = 0.012$). Lastly, human-written text had (just) significantly higher adequacy compared to the one-shot text, with a mean difference of − 0.29, 95% CI [− 0.57, − 0.001], $p = 0.048$. However, the non-significant results ($p > 0.05$) also reveal pertinent insights, as they indicate that AI-generated texts are not qualitatively distinguishable from human-written ones in certain aspects.

Hence, regarding correctness, there was no significant difference between any of the text types. In terms of coherence, there was no significant difference between

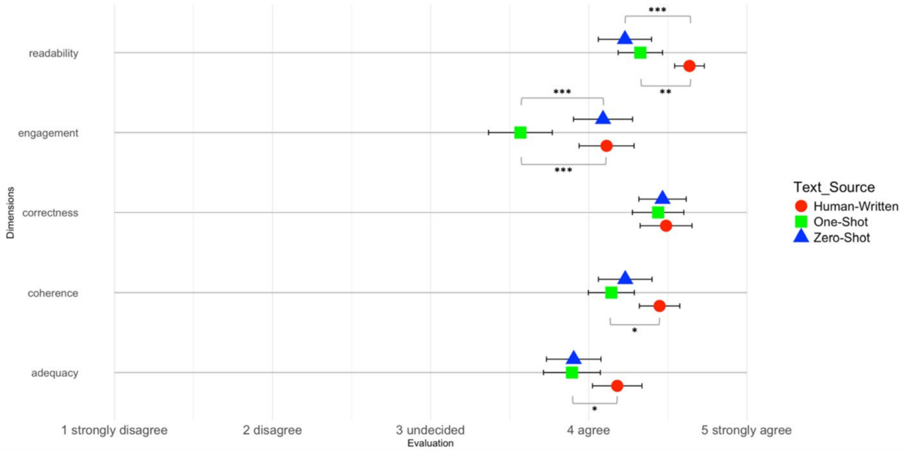

Fig. 2 Narrative text quality evaluation based on dimensions. With CI of 95% and $p < .05^*$, $p < .01^{**}$, $p < .001^{***}$

human-written and zero-shot. There was also no significant difference between the two AI-generated texts regarding readability, correctness, coherence, and adequacy.

Informative texts Similarly, a multivariate analysis of variance (MANOVA) was conducted to examine the effect of text types (human-written, one-shot, zero-shot) on the five quality dimensions as dependent variables. Using Pillai's Trace, there was a significant main effect of text type on the ratings of the quality dimensions, $V=0.077$, $F(10, 522)=2.093$, $p=0.023$. Precisely, post-hoc Tukey HSD results (see Fig. 3) based on the significant univariate ANOVA analyses revealed a significant difference in readability between human-written and one-shot, with one-shot being rated more highly (0.255, 95% CI [0.013, 0.496], $p=0.036$).

The mean difference between the two AI-generated texts was also significant, with one-shot performing better than zero-shot (− 0.258, 95% CI [− 0.500, − 0.017], $p=0.033$). In terms of coherence, there was a significant difference between zero-shot and one-shot (− 0.251, 95% CI [− 0.473, − 0.028], $p=0.023$), in favor of the one-shot text. The last significant effect was found for the texts' adequacy, as the one-shot text obtained higher results than the zero-shot text (− 0.444, 95% CI [− 0.759, − 0.129], $p=0.003$). The majority of non-significant results ($p>0.05$) between the text types indicates high resemblance between human-written and AI-generated texts across the quality dimensions.

Expert subgroup A further multivariate analysis of variance (MANOVA) was conducted to investigate whether there were substantial differences in the evaluation of experienced reviewers with specialized knowledge in teaching or test development ($n=34$) compared to the remaining sample ($n=55$). The MANOVA using Pillai's Trace revealed no significant main effect of group (general population vs. experts) on the quality dimensions and the text discernibility, neither for narrative ($V=0.083$, $F(16, 246)=1.39$, $p=0.16$) nor for informative texts ($V=0.072$, $F(16, 246)=1.20$, $p=0.27$).

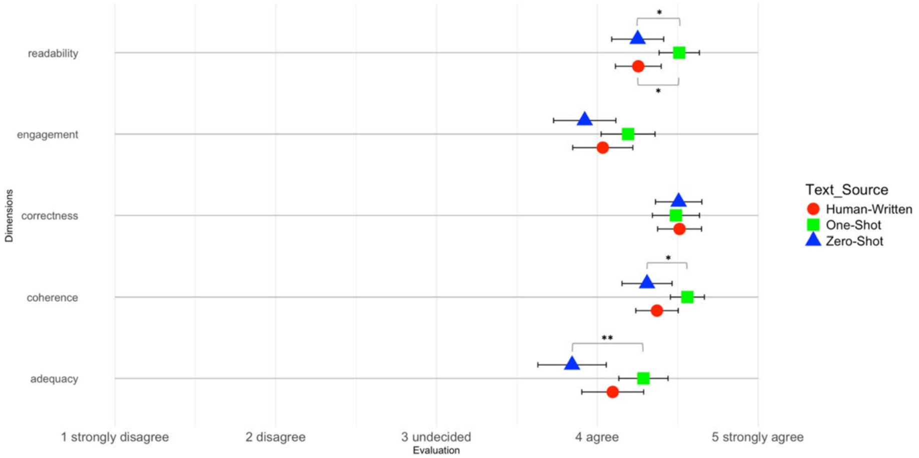

Fig. 3 Informative text quality evaluation based on dimensions. With CI of 95% and $p<.05$*, $p<.01$**, $p<.001$***

Table 3 Descriptive statistics for narrative texts

Scale	M	SD	Median	Min	Max	Q1	Q3	Cronbach's α
Readability	4.40	0.67	4.67	1.67	5	4	5	0.76
Engagement	3.92	0.92	4	1	5	3.5	4.5	0.72
Correctness	4.46	0.75	5	1	5	4	5	0.87
Coherence	4.27	0.71	4.5	1.83	5	4	4.83	0.89
Adequacy	3.99	0.81	4	1	5	3.5	4.5	0.82

Table 4 Descriptive statistics for informative texts

Scale	M	SD	Median	Min	Max	Q1	Q3	Cronbach's α
Readability	4.34	0.69	4.67	2	5	4	5	0.75
Engagement	4.05	0.87	4	1	5	3.5	5	0.74
Correctness	4.5	0.67	5	2	5	4	5	0.71
Coherence	4.41	0.63	4.67	2.33	5	4	5	0.89
Adequacy	4.07	0.91	4	1	5	3.5	5	0.87

Descriptive and assumption tests The descriptive results (see Tables 3 and 4) of the quality dimensions are characterized by high mean ratings with limited variability across the text types, indicating a ceiling effect. The moderate to high internal consistencies, evidenced by Cronbach's alpha values ranging from $\alpha = 0.71$ to $\alpha = 0.89$, demonstrate that the items within each dimension reliably measure the same underlying construct. The correlations between the dependent variables range from $r = 0.32$ to $r = 0.69$ for narrative and from $r = 0.37$ to $r = 0.78$ for informative text type, suggesting that while there is some degree of relationship between the dimensions, multicollinearity is not a significant concern. With regards to the MANOVA analyses, the Shapiro–Wilk test for multivariate normality indicated that the residuals for all scales significantly deviated from normality ($p < 0.001$) in both narrative and informative texts. Hence, the statistically significant differences found in some dimensions are likely marginal and not indicative of substantial perceptual differences among the participants. A possible reason could be that the text types likely have inherent differences in variability, due to how they are written and consequently, how they are perceived and rated by the reviewers. Still, the significant differences of the post-hoc results should be interpreted with some caution. The ceiling effect might result from using carefully pre-screened texts from a high-stakes assessment context, where both human- and AI-generated texts were either already validated or closely aligned with established design standards.

Discernibility of human-written vs. AI-generated texts The analysis of the discernibility between human-written and GPT-4-generated texts yielded a number of notable findings (see Fig. 4). In all text categories, a certain proportion of responses remained undecided, ranging from approximately one-sixth to one-fourth among all respondents for the distinct text types. It was found that none of the texts were correctly identified by an absolute majority of reviewers, with no text type achieving more than

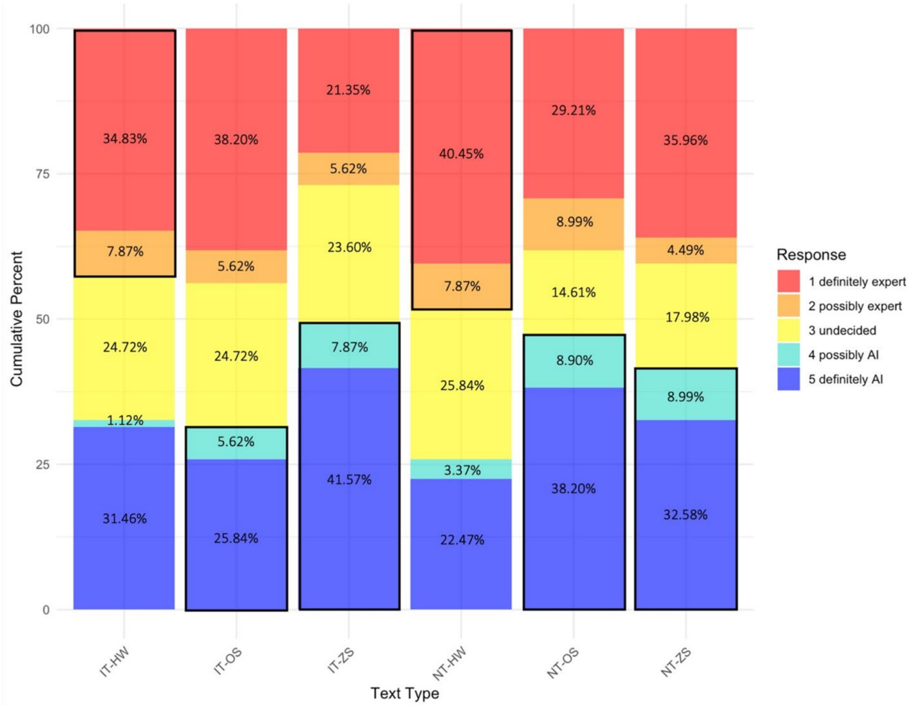

Fig. 4 Expert versus AI assessment by text type (cf. Casal & Kessler, 2023). *IT-HW* Informative Text Human-Written, *IT-OS* Informative Text One-Shot, *IT-ZS* Informative Text Zero-Shot, *NT-HW* Narrative Text Human-Written, *NT-OS* Narrative Text One-Shot, *NT-ZS* Narrative Text Zero-Shot. Black framings represent correct identification

50% of correct identification. The two texts that were correctly rated by the greatest proportion of participants were the informative zero-shot text (49.44%) and the narrative human-written text (48.32%). The mean accuracy in identifying narrative texts was slightly higher (45.67%) than that for informative texts (41.20%). More specifically, the informative human-written text demonstrated a notable proportion of participants who correctly classified it as definitely expert (34.83%) or possibly expert (7.87%), although the majority remained undecided (24.72%) or erroneously estimated the text to be authored by AI (32.58%). Regarding the informative one-shot text, only around one-third of the participants (31.46%) were able to correctly identify it as AI-written, while a considerable proportion estimated it to be human-written (43.82%) and almost one-fourth were undecided (24.72%). The informative zero-shot text was the most straightforward of the AI-generated texts to distinguish from human-written ones, with almost half of all reviewers (49.44%) correctly identifying it as AI. Here, the majority of participants across all text types indicated that they were definitely sure of their answer (41.57%).

In terms of the narrative human-written text, almost half of reviewers could detect the expert's authorship (48.32%), while the remaining half were equally uncertain (25.84%) or incorrectly estimated that AI (25.84%) had authored the text. Between the two human-written texts, the narrative style was correctly identified by a greater number of participants (48.32%) than the informative style (42.70%). In the narrative one-shot text, a considerable proportion correctly identified the text as having been artificially generated (47.10%), substantially more than for the informative one-shot text (31.46%). However, the majority remained undecided (14.61%) or incorrectly rated the text to be human-written (38.20%). The percentage of respondents who indicated that the text was definitely AI-generated was higher in comparison to the narrative zero-shot text (38.20% vs. 32.58%). Lastly, the narrative zero-shot text elicited an almost identical proportion of accurate responses regarding its author, with 41.57% correctly identifying the text as AI-generated and 40.45% incorrectly attributing it to a human author. The remainder of respondents were undecided (17.98%). It proved more difficult to correctly identify the narrative zero-shot text as AI than the informative zero-shot text (41.57% vs. 49.44%).

Discussion

The findings of this study offer valuable insights into the effectiveness of distinct prompt designs for generating reading comprehension texts using GPT-4, particularly regarding diverse aspects of their quality and their overall suitability for educational assessment. In general, the six texts administered were of similarly high quality, making it difficult for the reviewers to accurately distinguish between those written by humans and those generated by GPT-4.

RQ2: Quality and adequacy for educational assessment In terms of narrative style, the human-written text ("Little Monkey") demonstrated the highest quality across all dimensions, while the two AI-generated texts ("Treasure Hunt" and "First day of school") exhibited similar results except in the dimension of engagement. In this instance, both human-written and zero-shot texts were found to be significantly more suitable and engaging for questions, as well as to be written in a more lively and interesting way than the one-shot text. This finding aligns with previous research conducted by Bezirhan and von Davier (2023), who stated that their generated literary stories were less distracting and more engaging than the generated informational passages. As the aspects of creativity and flexibility are more pertinent to perceive a narrative story as engaging, zero-shot prompting could facilitate greater adaptability and originality. It affirms Reynolds' and McDonell's (2021) assertion that the zero-shot format is capable of achieving comparable and even superior results to those obtained through few-shot prompting. In contrast, the one-shot text appeared to be too restricted in adapting the structural and linguistic constraints of the text model and the reference passage, resulting in a compromise of its engaging quality.

Furthermore, human-written texts were found to be significantly more coherent than AI-generated texts, indicating that humans may still outperform GPT-4 in maintaining thematic and logical consistency. Therefore, the original texts were deemed to be more suitable for assessing 5th grade reading comprehension in national school monitoring by the participants. With regard to the texts' correctness, both prompt engineering approaches were able to produce grammatically and orthographically correct texts, thereby demonstrating the considerable linguistic capabilities of GPT-4 in the German language (OpenAI, 2023). This strength was observed for both narrative and informative texts.

Interestingly, the one-shot text ("Earthworms") performed best across most quality dimensions for informative texts, indicating that GPT-4 is capable of effectively learning and replicating the format and style from a few instructions and examples. This finding is consistent with previous research, in that clear instructions may enhance the performance of GPT models. Additionally, one-shot learning, including specific contextual information about the target group, may be most effective in matching the topic and difficulty of the original text (Bezirhan & von Davier; OpenAI, 2023). The relatively straightforward structure of the informative reading assessment, such as presenting facts about animals, appears to be more readily replicable by the model than the more nuanced writing typical in narrative texts. This is particularly evident in the readability and coherence dimension, where one-shot prompting resulted in significantly higher outcomes. It is likely that the positive evaluations in the majority of the quality dimensions influenced the rating of the one-shot text as being significantly more suitable for reading comprehension in national school monitoring, than, for instance, human-written texts. This finding is supported by a prior study (Xiao et al., 2023), wherein evaluators similarly deemed ChatGPT-generated passages highly suitable for reading comprehension exercises, often rating them more highly than current educational materials.

In conclusion, the results indicate that one-shot prompting is highly effective for the creation of high-quality informative texts with GPT-4, whereas human-written texts remain superior in narrative contexts. Zero-shot generation, due to its enhanced flexibility and creativity, may offer new thematic content suggestions, although still requiring refinement by human item developers.

RQ3: Capacity to distinguish between humanly authored and AI-generated text passages The considerable proportion of undecided responses across all text types highlights the increasingly sophisticated capabilities of AI in text generation**.** The fact that none of the texts were correctly identified by an absolute majority of reviewers aligns with prior research (Casal & Kessler, 2023) and implies that both human and AI-generated texts comprise elements that can mislead evaluators. This indicates that AI is capable of producing content that closely mimics human writing across various genres, particularly informative texts, as reflected in a lower correct identification percentage. Moreover, the higher correct identification rates for the informative zero-shot text and the narrative human-written text suggest that certain styles or content types may be more easily recognized as either human or AI-generated. For instance, informative texts may exhibit more structured and predictable patterns that AI can replicate effectively, while narrative texts may display more nuanced elements of creativity and imagination. The "definitely AI" identification is particularly high for the zero-shot texts, indicating that these texts may exhibit characteristics that reviewers associate with artificial generation. This seems logical, given that no original text passage was provided as reference in this prompt design, which has the effect of enhancing GPT-4's flexibility but not really adopting a human writing style. The higher rates of "possibly expert" and "definitely expert" for one-shot texts compared to zero-shot texts indicate that providing a reference or example in the prompt improves the human-likeness of the generated text. This aligns with Xiao et al. (2023), who found that the effect of human-likeness was more pronounced in one-shot setting, as the model effectively imitates the style and structure of the provided reference passage.

Similarly to the results of the dimensional quality assessments, the provision of reference texts in one-shot prompting enhances human-likeness but may constrain creativity. This not only affects the perceived engagement of the generated text but also its discernibility. In general, this comparison underscores the potential of GPT-4 to mimic human writing, hence presenting considerable potential for contributing to the item development process in educational assessments.

General discussion

The present study aimed to address gaps in the current research landscape regarding the generation of reading comprehension texts with GPT-4. It also emphasized the involvement of SMEs throughout the text development process, in order to maintain high-quality standards. In light of the specific context in Luxembourg, the study sought to explore the potential generalizability of LLM technology to languages other than English, precisely German, in the development of reading comprehension texts. The primary objectives were to determine whether GPT-4 could enhance human task performance, to identify the most suitable prompt engineering approach, and to evaluate the overall quality of generated texts.

The outcomes of the focus group discussion (study 1) indicated the potential value of artificial intelligence (AI), such as GPT-4, as an assistive tool in the initial phase of text creation. In the multilingual context of Luxembourg, item developers encounter difficulties in sourcing language-appropriate texts from the literature. Hence, this approach can efficiently provide first drafts, which can then undergo further refinements by SMEs. This idea was corroborated by the quantitative findings (study 2), demonstrating that the quality of AI-generated texts is either comparable to or exceeds that of human-written texts. It is important to note that the differences observed in certain dimensions are only marginal and that reviewers were unable to consistently identify authorship origins. This further substantiates the conclusion that GPT-4 is capable of emulating human-written texts with respect to both linguistic style and overall quality.

Study 2 evaluated different prompt engineering approaches and their impact on output quality. One-shot prompting, which entails providing GPT-4 with an original text to orient its response, proved to be highly effective in producing informative texts, but suffered creative trade-offs in narrative texts. This is consistent with the interviewees' (study 1) concern that AI merely replicates learned patterns and therefore cannot approximate the creativity inherent in the human mind. Still, outputs generated via one-shot prompting were more likely to be misidentified as human, as the presence of a reference text increased human-likeness. Zero-shot prompting demonstrated greater flexibility and creativity, although higher correct identification rates suggest that specific text characteristics are associated with artificial generation by reviewers. With regard to the practical use of LLMs for language assessment in ÉpStan, or test development in general, the results allow for several recommendations: Zero-shot prompting could be employed for generating new ideas, while one-shot prompting could be beneficial for repurposing "old" and adapting existing texts, or when suitable texts are difficult to source.

Concerning the slightly higher ratings of human-written narrative texts, it is worth mentioning that narrative engagement and coherence rely on fluid emotional transportation and contextual adaptation, which human storytelling may intuitively achieve through lived experiences. In contrast, LLMs may exhibit a preference for conceptual stability and for nouns, adjectives and prepositions, whereas human retellings of stories are notably more creative and characterized by a preference for verbs, adverbs and pronouns (Breithaupt et al., 2024). GPT-4 operates on the basis of statistical pattern recognition rather than conscious thought, emotion or personal experience. While this pattern-based learning allows for highly structured and informative text generation, it may limit the adaptive refinement capabilities inherent in human narrative writing. Recent advancements like GPT-4o aim to improve deliberate reasoning (OpenAI, 2024), although GPT-4's limitations in coherence and engagement for narrative texts could also be explained by the use of singular rather than repeated instructions in the present study. Future studies could explore methods for improving narrative engagement in AI-generated texts, including (1) adjusting generation parameters such as temperature or top-k sampling to increase lexical and structural variability; (2) using multi-step or chained prompting strategies (e.g., few-shot examples or structured scaffolds) to support coherent and emotionally rich storytelling; and (3) integrating human-in-the-loop feedback to refine affective nuance and narrative depth. Additionally, a clearer operationalization of narrative engagement in this context could guide more targeted development and evaluation efforts.

The incorporation of a model-based framework, such as the TACM, provided a structured foundation for aligning the generated texts with educational standards and cognitive task demands. In our case, this ensured consistency with existing test design practices in the ÉpStan and supported interpretive validity in terms of construct representation. In practice, this approach allows for the derivation of specific prompt templates (i.e. in the form of a CustomGPT model) according to a theoretically founded methodology that could be continuously updated by item developers. The ability of LLMs to generate a multitude of text passages that fulfil the same criteria could be employed in the design of parallel test forms (even across languages) and most importantly, address the growing demand for novel test content. Although the TACM proved valuable in our study, alternative model-based frameworks should be explored in future works to distinguish an optimal approach.

Several limitations of the study are noteworthy. With regard to the focus group discussion, it is possible that bias may have been introduced due to group dynamics. The limited number of participants in both the qualitative and quantitative evaluations, as well as the administration of only four generated texts in the evaluation questionnaire, may limit the generalizability of the results. On the grounds that the sample size in the present study was greater than 30 (Central Limit Theorem, Field et al., 2012), and constituted similar group sizes, a certain robustness of the MANOVA analyses towards the violation of multivariate distribution was assumed (Tabachnick & Fidell, 2013). However, the inherent properties of the data (ceiling effects, skewness) must be considered when interpreting the results. We acknowledge that the high overall ratings of the texts may reflect limited score variance, and future research could address this by employing more sensitive rating scales, anchored evaluation procedures, or alternative approaches such as comparative judgment to enhance differentiation.

Legal and ethical challenges associated with AI-generated content must also be acknowledged—especially regarding authorship, copyright, and potential plagiarism.

In the European context, the current legal status of texts produced by large language models remains ambiguous, particularly given the opacity of training corpora and the lack of clear attribution. Additionally, the use of AI in multilingual assessment contexts introduces further complexity, as cultural or pragmatic biases embedded in training data may affect language-specific outputs in subtle but important ways. These concerns underline the need for a human-in-the-loop approach in test development and for institutional frameworks that address issues of data transparency, accountability, and intellectual property. Future regulatory developments may clarify these questions, at which point empirical evaluation of psychometric properties and best practice guidelines will become essential. In such a scenario, the psychometric properties could be empirically investigated by utilizing them in educational large-scale assessments, such as the ÉpStan, and guidelines for best practices could be provided. Future research should investigate the quality of GPT-generated reading comprehension texts in a variety of languages, with a view to establishing their efficacy in cross-linguistic applications. Given the unique multilingual setting of this study, future research should explore the generalizability of AI-generated test content across languages and cultural contexts. This includes investigating how linguistic subtleties and culturally embedded meanings may influence interpretation and perceived quality. Such work could help identify and mitigate potential cultural biases, reinforcing the need for continued human oversight in multilingual item development.

While different prompt designs were explored in the present study, future studies could collaborate with prompt engineers and experts from the field of artificial intelligence to exploit the potential of this technology in a more comprehensive way. By the time of completing this study, an even more advanced and human-like iteration called GPT-4o ("omni") has been released (OpenAI, 2024). Future studies are needed to replicate these findings and further investigate LLM's creativity considering the newest models for text generation, in that increasingly enhanced performance is assumed.

Conclusion

Through a mixed approach combining qualitative and quantitative methodologies, the present study highlights the significant potential of using GPT-4 as an assistive tool in the flexible and efficient development of high-quality text passages for reading comprehension assessment. This is of particular relevance in multilingual educational contexts such as Luxembourg, where conventional test development is associated with substantial resource costs. Despite the potential of AI, item developers must remain responsible for thoroughly verifying, and refining the generated texts, to ensure they are correct, align with the intended requirements, and are free from biases. It can be reasonably assumed that the augmented intelligence approach, which effectively combines innovative technologies with human expertise, will become a dominant force in the future. This will undoubtedly lead to the emergence of a multitude of new possibilities, with the present study being only one example. All while having ethical implications in mind, "[w]e are entering a new paradigm of human–computer interaction in which anyone who is fluent in natural language can be a programmer" (Reynolds & McDonnell, 2021, p. 8).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-025-00255-w.

Additional file 1

Additional file 2

Additional file 3

Acknowledgements

The authors want to thank the language test developers at the Luxembourg Centre for Educational Testing for participating in the focus group interview. A special thank you goes to Katharina Antonia Michiko Tremmel for providing the human-written texts and for giving extensive feedback on reading comprehension text characteristics. The authors want to additionally thank all participants reading and evaluating the AI-generated and human-written texts.

Author contributions

LRYS planned the study, created the materials, conducted the focus group interview and the comparative study. LRYS analyzed the data and wrote major parts of the manuscript. PS planned the study, oversaw and reviewed the study's materials and focus group guidelines. PS conceived of and supervised the data analysis and contributed to writing the manuscript. Both authors read and approved the final manuscript.

Funding

Not applicable.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Data protection and ethical issues were handled according to the guidelines provided by the Luxembourg Commission nationale pour la protection des données (CNPD) [National Commission for Data Protection], the University of Luxembourg and the American Psychological Association. Participation was voluntary and could be withdrawn at any moment of the data collection (focus group and online survey alike).

Consent for publication

No parts of the enclosed manuscript have been previously published or are currently under review in any other journal. Likewise, the full report is not under consideration for publication elsewhere, and there is no overlap with other papers.

Competing interests

The authors declare that they have no competing interests.

Received: 27 July 2024 Accepted: 5 July 2025
Published online: 17 July 2025

References

Ackerman, R., & Balyan, R. (2023). *Automatic Multilingual question generation for health data using LLMs*. Springer Nature Singapore.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., Von Davier, A. A., & Duolingo. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2022.903077

Bezirhan, U., & Von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education. Artificial Intelligence, 5*, Article 100161. https://doi.org/10.1016/j.caeai.2023.100161

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler*. Springer.

Breen, R. L. (2006). A practical guide to focus-group research. *Journal of Geography in Higher Education, 30*(3), 463–475. https://doi.org/10.1080/03098260600927575

Breithaupt, F., Otenen, E., Wright, D. R., Kruschke, J. K., Li, Y., & Tan, Y. (2024). Humans create more novelty than ChatGPT when asked to retell a story. *Scientific Reports, 14*(1), 875.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arxiv.2005.14165

Bulut, O., & Yildirim-Erbasli, S. N. (2022). Automatic story and item generation for reading comprehension assessments with transformers. *International Journal Of Assessment Tools in Education*, *9*(Special Issue), 72–87. https://doi.org/10.21449/ijate.1124382.

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3), Article 100068. https://doi.org/10.1016/j.rmal.2023.100068

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *arXiv*. https://doi.org/10.48550/arxiv.2006.14799

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*. https://doi.org/10.3389/feduc.2023.858273

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE.

Gierl, M. J., & Haladyna, T. M. (2013). *Automatic item generation: Theory and Practice*. Routledge.

Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education, 47*(7), 726–733. https://doi.org/10.1111/medu.12202

Gierl, M. J., Lai, H., & Matovinovic, D. (2020). *Augmented intelligence and the future of item development*. Information Age Publishing Inc.

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.

LimeSurvey GmbH. (n.d.). *LimeSurvey: An Open-Source survey tool.* LimeSurvey GmbH. http://www.limesurvey.org

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. *Routledge eBooks*. https://doi.org/10.4324/9780203850381

Howitt, D., & Cramer, D. (2017). *Introduction to research methods in Psychology*. Pearson Education.

Jung, J. Y., Tyack, L., & Von Davier, M. (2024). Combining machine translation and automated scoring in international large-scale assessments. *Large-Scale Assessments in Education*. https://doi.org/10.1186/s40536-024-00199-7

Keyser, J. (2023). *Jon (Jody) Keyser on LinkedIn: Fine-tuning vs. Prompt Engineering*. https://www.linkedin.com/posts/jodykeyser_fine-tuning-vs-prompt-engineering-activity-7135835527051677696-0Zdh

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2018). A Cost-Benefit analysis of automatic item generation. *Educational Measurement, 38*(1), 48–53. https://doi.org/10.1111/emip.12237

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education. Artificial Intelligence, 6*, Article 100210. https://doi.org/10.1016/j.caeai.2024.100210

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*. https://doi.org/10.1007/s10639-023-12249-8

Leigh, D. (2009). SWOT analysis. *International Society for Performance Improvement*. https://doi.org/10.1002/9780470592663.ch24

Levitina, A. (2024). *Generative AI and Intellectual Property Rights in the EU context - Logan & Partners*. Logan & Partners. https://www.loganpartners.com/generative-ai-and-intellectual-property-rights-in-the-eu-context/

Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*. https://doi.org/10.1016/j.system.2024.103344

Lui, A., & Lamb, G. W. (2018). Artificial intelligence and augmented intelligence collaboration: Regaining trust and confidence in the financial sector. *Information & Communications Technology Law, 27*(3), 267–283. https://doi.org/10.1080/13600834.2018.1488659

Mayring, P. (2003). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Beltz.

Men.lu. (2023). *Die Sprachen in der Luxemburger Schule.* Site Du Ministère De L'Éducation Nationale, De L'Enfance Et De La Jeunesse. https://men.public.lu/de/themes-transversaux/langues-ecole-luxembourgeoise.html

OpenAI. (n.d.). *Prompt Engineering*. OpenAI Platform. https://platform.openai.com/docs/guides/prompt-engineering

OpenAI. (2023). GPT-4 Technical Report. In *arXiv: Vol. 2303.08774v6* [Technical report]. https://arxiv.org/pdf/2303.08774.pdf

OpenAI. (2024). *Hello GPT-4o: We're announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time.* openai.com. https://openai.com/index/hello-gpt-4o/

Pankiewicz, M., & Baker, R. S. (2023). Large Language Models (GPT) for automating feedback on programming assignments. *arXiv*. https://doi.org/10.48550/arxiv.2307.00150

European Parliament. (2024). *Artificial Intelligence Act*. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the Few-Shot paradigm. *arXiv*. https://doi.org/10.48550/arxiv.2102.07350

RStudio Team (2020). *RStudio: Integrated Development for R.* RStudio. http://www.rstudio.com/

Säuberli, A., & Clematide, S. (2024). Automatic generation and evaluation of reading comprehension test items with large language models. *arXiv*. https://doi.org/10.48550/arxiv.2404.07720

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. In *Educational Measurement: Issues and Practice* (pp. 5–18).

Schreiner, M. (2023). GPT-4 architecture, datasets, costs and more leaked. *THE DECODER*. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/#google_vignette

Épreuves Standardisées. (n.d.). EpStan General Information. https://epstan.lu/en/general-information/

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics.* Pearson.

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. J. (2024). A Review of Automatic Item Generation Techniques Leveraging Large Language Models. [Preprint]. https://doi.org/10.35542/osf.io/6d8tj

Tomikawa, Y., Uto, M. (2024). Difficulty-controllable reading comprehension question generation considering the difficulty of reading passages. *International Conference On Computers in Education*. https://doi.org/10.58459/icce.2024.4931.

Ugen, S., Schiltz, C., Fischbach, A., & Cate, I. P. (2021). *Lernstörungen im multilingualen Kontext: Diagnose und Hilfestellungen.* https://doi.org/10.26298/bw1j-9202

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv*. https://doi.org/10.48550/arxiv.1706.03762

Von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Wang, Z., Valdez, J., Mallick, D. B., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-031-11644-5_13

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications. *Conference: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. 2023 https://doi.org/10.18653/v1/2023.bea-1.52

Yang, S. J. H., Ogata, H., & Matsui, T. (2023). Guest Editorial: Human-centered AI in Education: Augment Human Intelligence with Machine Intelligence. *Educational Technology & Society, 26*(1), 95–98.

Zheng, N., Liu, Z., Ren, P., Ma, Y., Chen, S., Yu, S., Xue, J., Chen, B., & Wang, F. (2017). Hybrid-augmented intelligence: Collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering/frontiers of Information Technology & Electronic Engineering, 18*(2), 153–179. https://doi.org/10.1631/fitee.1700053

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P. F., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv*. https://doi.org/10.48550/arxiv.1909.08593

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.