Journal of Artificial Intelligence and Computing Applications (2026) 4(1): 31–39

MAIKRON

Journal of Artificial Intelligence and Computing Applications

Review article

Automated multiple-choice question generation for soft skills development using LLMs: a short narrative review

Luis Cardena-Ley[1]

[1]Tecnológico Nacional de México/IT de Mérida, Yucatán, México

ABSTRACT

The development of soft skills is crucial in higher education and professional training, yet traditional assessment methods often struggle to provide personalized and scalable solutions. This narrative review explores the potential of Large Language Models (LLMs), specifically GPT-based models, in generating multiple-choice questions (MCQs) aimed at soft skills development. A structured analysis of recent adjacent literature suggests that LLMs may offer considerable promise for automating the creation of diverse, contextually relevant, and construct-specific MCQs across various domains, including psychological assessments, medical training, and educational settings. Key techniques such as prompt engineering, model fine-tuning, and retrieval-augmented generation are identified as effective methods for enhancing the accuracy and relevance of AI-generated questions. However, challenges such as biases, factual inaccuracies, and the need for human oversight remain significant concerns. The reviewed studies consistently highlight the importance of integrating human expertise to ensure the validity, fairness, and quality of the generated content. Despite these challenges, the use of LLMs could streamline the creation of personalized learning tools, offering scalable solutions for soft skills assessment. Future research should focus on refining AI techniques, developing robust evaluation frameworks, and exploring the impact of AI-generated MCQs on learners' skill development. This review emphasizes the need for a hybrid approach, combining AI's efficiency with expert validation, to fully realize the potential of LLMs in soft skills education.

Keywords: large language models, multiple-choice questions, soft skills

1. Introduction

The development of soft skills, such as communication, critical thinking, teamwork, and problem-solving, has become increasingly crucial in today's fast-paced and dynamic workforce [1]. Unlike hard skills, which are specific, teachable abilities or knowledge, soft skills are more intangible and harder to quantify [2]. However, they are essential for effective collaboration, leadership, and adaptability in various professional settings [2]. As organizations recognize the value of these competencies, there is a growing demand for innovative methods to nurture and assess them [3]. One promising approach is the use of multiple-choice questions (MCQs), which can be strategically designed to promote reflection and learning [4]. With advancements in artificial intelligence (AI), particularly in Large Language Models (LLMs), there is potential to generate these MCQs using automatic item generation (AIG), creating personalized and effective tools for soft skills development [5].

Despite the recognized importance of soft skills, traditional methods of developing and assessing these competencies often fall short in both engagement and effectiveness [6]. Existing literature highlights several ap-

E-mail address: le15080882@merida.tecnm.mx
https://doi.org/10.5281/zenodo.20802609
 ISSN: 3061-8843

proaches to soft skills development, including experiential learning, role-playing, and self-assessment [4], yet there remains a persistent challenge in creating scalable and personalized learning experiences [5]. Foundational work has emphasized the necessity of integrating emotional intelligence and interpersonal skills into professional training programs [3]. However, these approaches often rely on human-mediated interventions, which can be resource-intensive and may lack consistency [2]. More recent research has explored the potential of using technology-based solutions, such as e-learning platforms and virtual simulations [4], but these too have limitations in adaptability and real-time feedback [7]. Notably, there is a limited body of research examining the automatic MCQ generation specifically aimed at soft skills development, particularly leveraging advanced AI models like LLMs [5]. This gap presents an opportunity to explore innovative approaches that can bridge the existing limitations and provide scalable, personalized solutions for soft skills acquisition [4].

The objective of this Short Narrative Review (SNR) is to explore the potential of leveraging artificial intelligence, specifically LLMs, to automatically generate MCQs that facilitate the development of soft skills, hereafter referred to as soft skills MCQs. By addressing the identified gaps in the existing literature, this review aims to provide insights into the capabilities and limitations of LLMs in creating personalized, engaging, and effective soft skills MCQs. The guiding research question for this review is: "Can LLMs be utilized to automatically generate MCQs that help individuals develop specific soft skills?" Through a comprehensive analysis of current AI methodologies, practical applications, and case studies, this review seeks to uncover innovative approaches that can bridge the existing limitations in soft skills development and offer scalable, technology-driven solutions.

Given the vast landscape of research on soft skills, AI, and assessment, this review adopts a focused approach, prioritizing the analysis of the most critical and high-quality contributions to the field. To ensure a rigorous and impactful synthesis, the selection process involved a targeted search using Google Scholar, followed by a meticulous screening of results. Only original research articles indexed in the Journal Citation Reports (JCR) and belonging to the top two quartiles (Q1 and Q2) were included. This deliberate strategy allows for a concentrated examination of influential and peer-validated studies, enabling a deeper and more insightful appraisal of the current state of knowledge regarding the automatic generation of soft skills assessment tools using LLMs.

This review systematically explores the potential of leveraging LLMs for the automated generation of soft skills MCQs. Section 2, *Methodology*, outlines the literature review process, detailing the databases searched, keywords utilized, and selection criteria applied to ensure a transparent, replicable, and credible approach. Section 3, *Thematic Overview*, synthesizes the main findings, identifying emergent themes, trends, and patterns in the use of LLMs for MCQ generation, while critically analyzing their applicability to soft skills development and highlighting gaps in the existing literature. Section 4, *Discussion*, interprets these findings, examining their implications for the use of AI in educational contexts, particularly in designing soft skills MCQs, and addresses limitations through comparative analysis. Finally, Section 5, *Conclusion*, consolidates the key insights, underscores their significance for AI-driven educational tools, and proposes future research directions to advance the development of LLM-generated soft skills MCQs. Through this structured approach, the review aims to contribute to the growing discourse on AI's role in enhancing educational methodologies for skills development.

2. Methodology

The methodological framework for this review follows a structured and transparent approach to synthesize recent and relevant research in the field. While this is not a systematic review, the study selection process aligns with key principles of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology to ensure clarity, rigor, and replicability [8]. PRISMA guidelines provide a structured approach to literature identification, screening, and inclusion, which enhances the reliability of narrative reviews by minimizing selection bias [8]. Given the SNR format, the focus is not on comprehensive coverage but on curating a targeted subset of impactful studies that provide valuable insights into the research topic [9]. To achieve this, a systematic selection process was implemented, emphasizing recent publications from high-quality, peer-reviewed sources while maintaining transparency in inclusion and exclusion criteria. This methodological rigor ensures that the findings are based on a well-defined and reproducible literature selection process, strengthening the credibility of the review [9].

The identification phase involved a structured search strategy to retrieve relevant studies on the research topic. The search was conducted using Google Scholar, a widely used academic search engine that provides broad access to peer-reviewed journal articles across various disciplines [10]. This platform was selected for its comprehensive coverage of scholarly literature and its ability to index recent and high-impact research in artificial intelligence and computing [11]. A predefined search query, "intitle:("soft skills" OR "social skills" OR "21st century skills" OR "interpersonal skills" OR "skills" OR "skill") AND intitle:(generation) AND intitle:("multiple choice questions" OR "Multiple-choice questions" OR "automated item generation" OR "MCQs" OR "AIG") AND intitle:(LLM OR GPT OR LLMs)", was applied to ensure consistency in retrieval. To maintain relevance, the search was restricted to studies published from 2020 onward, thereby focusing on recent advancements in the field. This process yielded a total of 184 records, which were subsequently assessed through a multi-stage selection process to ensure that

only high-quality and thematically relevant studies were included in the review.

The screening phase aimed to refine the dataset by eliminating studies that did not meet the relevance and quality criteria [8]. The initial filtering process was conducted through title and abstract screening, where studies that were unrelated to the research topic were excluded. Additionally, works not published in peer-reviewed scientific journals were removed to ensure that only rigorously vetted research was considered. Review articles were also excluded to maintain a focus on primary research studies that present original findings rather than secondary analyses. As a result of this screening process, 153 studies were excluded, leaving 31 studies for further evaluation in the eligibility phase. This step ensured that the remaining literature was directly relevant, high-quality, and methodologically sound, aligning with the objectives of this Short Narrative Review [9].

The eligibility phase involved a comprehensive full-text assessment of the remaining studies to ensure their alignment with the review's objectives [12]. This stage applied stricter exclusion criteria to further refine the selection [13]. Studies were excluded if they were not written in English, as language consistency is essential for accurate interpretation and synthesis [14]. Additionally, articles that were outside the defined scope of the review were removed to maintain thematic relevance [15]. To ensure scientific rigor, only studies published in journals indexed in the JCR within the first two quartiles were retained [16]. Furthermore, studies with fewer than nine citations per year were excluded to prioritize research with established impact. As a result of this assessment, 23 studies were excluded, leaving a final selection of 8 studies for inclusion. These additional filters ensured that the final dataset consisted of highly relevant, peer-reviewed, and impactful research, strengthening the validity of the review's findings [17].

Following the multi-stage selection process, a final set of 8 studies was identified for inclusion in this review. These studies represent recent, high-impact, and thematically relevant contributions to the field, ensuring that the review reflects the latest advancements. An overview of the selected articles, including their titles, journal names, and publication years, is provided in Table 1. To enhance transparency, the study selection process is visually summarized in Figure 1 (PRISMA diagram), detailing the number of records screened, excluded, and retained at each stage. While this methodology ensures rigor and relevance, certain limitations must be acknowledged. The exclusive use of Google Scholar may have restricted the comprehensiveness of the search, as other databases could yield additional relevant studies [10]. Furthermore, the focus on recent publications may have excluded historically significant research, and the exclusion of lower-quartile or non-indexed journals might have omitted valuable contributions from emerging scholars or specialized domains [18]. Additionally, the citation-based filtering may introduce a bias favoring well-established studies while potentially overlooking influential newer research [19]. Despite these limitations, the structured selection process ensures a transparent, replicable, and methodologically rigorous approach to identifying the most impactful literature for this review [12].

3. Thematic Overview

This section organizes and synthesizes the key findings from the selected literature, grouping them into themes that reflect the major topics, trends, and areas of consensus or debate within the field. The focus is on the use of LLMs such as different versions of GPT to automatically generate soft skills MCQs. The selected articles demonstrate the potential of LLMs to transform AIG across various domains, including psychological assessments, education, and medical training. By structuring the discussion around these themes, we aim to provide a cohesive narrative that highlights significant contributions, contradictions, and gaps in the literature.

The use of LLMs for automatic MCQ generation has been explored in multiple contexts. For instance, [20] demonstrates how GPT-3 can generate personality test items with psychometric properties comparable to human-authored items. Similarly, [21] introduces the Psychometric Item Generator (PIG), which uses GPT-2 to create diverse and high-quality psychological scale items. These studies highlight the scalability and efficiency of LLMs in generating diverse and contextually relevant questions. The implementation of prompt techniques on large language models can facilitate the generation of soft skills MCQs. This is supported by findings from [23], which demonstrate that prompt engineering significantly enhances the quality of AI-generated questions.

The psychometric properties of AI-generated MCQs have been rigorously evaluated in several studies. For example, [22] demonstrates that fine-tuned GPT-2 models can generate construct-specific personality test items with high psychometric quality. Similarly, [24] evaluates the difficulty and discrimination levels of ChatGPT-generated medical questions, showing that they meet acceptable psychometric standards. These findings underscore the importance of construct-specificity in generating MCQs that align with specific soft skills, such as communication, teamwork, and problem-solving. The role of fine-tuning models and prompt engineering is critical for ensuring that generated questions are both relevant and effective for soft skills development.

The practical implications of using LLMs to generate soft skills MCQs are significant. For instance, [25] provides actionable guidance for educators to use ChatGPT in generating high-quality medical questions, reducing their workload and enabling personalized learning. Similarly, [26] demonstrates how LLMs can automate the creation of reading comprehension assessments, saving time and resources. However, challenges such as potential biases, factual inaccuracies, and the need for human oversight are consistently highlighted across studies. For example, [27] emphasizes the impor-

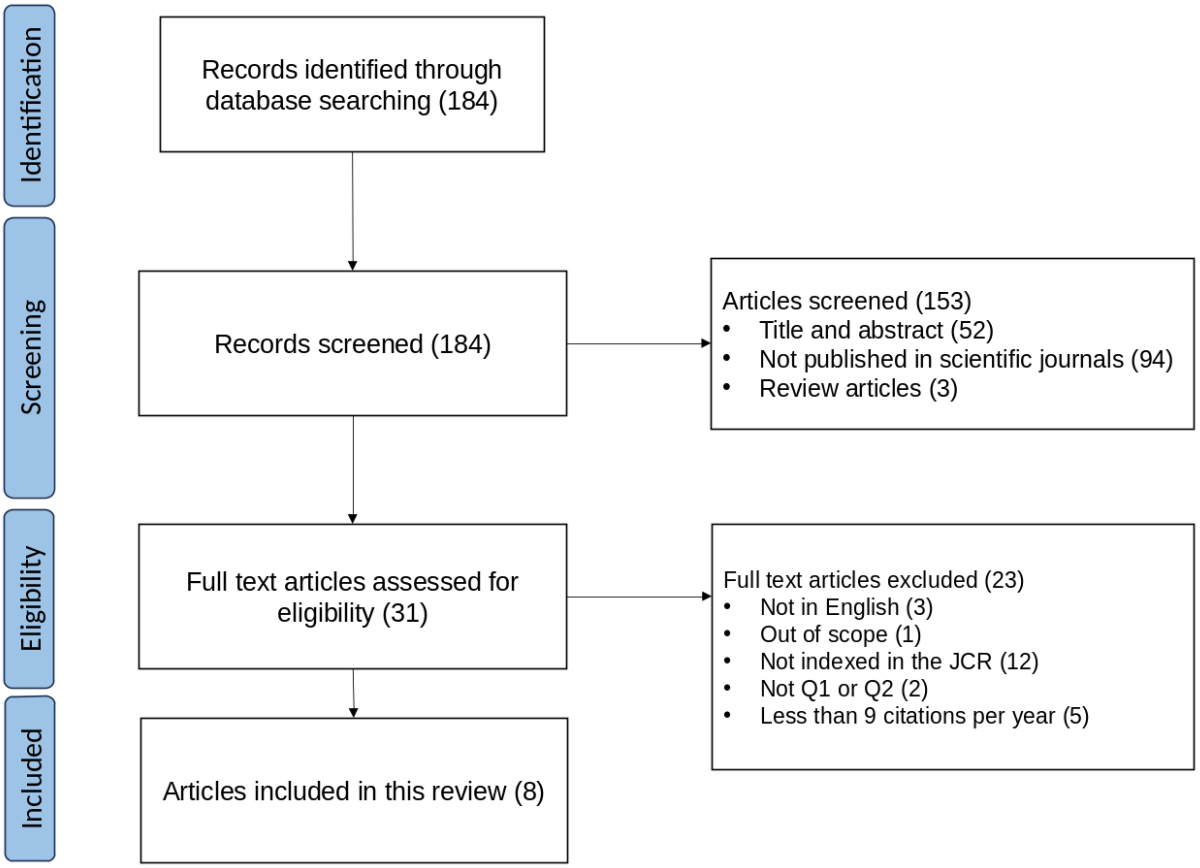

Figure 1. Flowchart of the literature identification, screening, eligibility assessment, and inclusion process.

Table 1. The final list of articles used in this review, including information for title, journal, year of publication and citation.

Title	Journal	Year	Citation
A Paradigm Shift from "Human Writing" to "Machine Generation" in Personality Test Development: an Application of State-of-the-Art Natural Language Processing	Journal of Business and Psychology	2023	[20]
Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development	Psychological Methods	2023	[21]
Transformer-based deep neural language modeling for construct-specific automatic item generation	Psychometrika	2022	[22]
Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education	Education and Information Technologies	2023	[23]
ChatGPT for generating multiple-choice questions: evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam	European Journal of Clinical Pharmacology	2024	[24]
Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT	Medical Teacher	2024	[25]
The interactive reading task: Transformer-based automatic item generation	Frontiers in Artificial Intelligence	2022	[26]
Biomedical knowledge graph-optimized prompt generation for large language models	Bioinformatics	2024	[27]

tance of integrating knowledge graphs with LLMs to improve accuracy and reduce biases in generated content.

The importance of AI-human collaboration is a recurring theme in the literature. The need for educators to guide LLMs in generating high-quality questions is highlighted in [25], ensuring scientific validity and contextual relevance. Similarly, [26] emphasizes the role of human review in ensuring fairness and bias-free content in AI-generated reading comprehension tasks. Ethical considerations, such as ensuring fairness, avoiding biases, and maintaining transparency in AI-generated content, are also critical. For instance, [27] discusses the ethical implications of relying on LLMs for generating biomedical text and the need for domain-specific knowledge to enhance accuracy.

The themes discussed in this section collectively demonstrate the potential of LLMs to generate high-quality soft skills MCQs. Areas of consensus include the scalability and efficiency of LLMs, the importance of construct-specificity, and the value of AI-human collaboration. However, ongoing debates about biases, factual inaccuracies, and the need for human oversight highlight the challenges that must be addressed to fully realize the potential of AI in this domain. These findings set the stage for the next sections of the review, which will explore the implications of these insights for soft skills development and future research directions. By leveraging the strengths of LLMs while addressing their limitations, we can create innovative and effective tools for soft skills MCQ generation.

4. Discussion

The reviewed literature indicates a significant potential for LLMs, including various GPT versions and ChatGPT, to enhance the efficiency of AIG across diverse assessment domains such as personality testing, psychological scales, medical examinations, and language education [20, 21, 22, 23, 24, 26, 25]. Key AI techniques identified include prompt engineering, ranging from few-shot learning to structured protocols and practical tips [20, 23, 25], and the fine-tuning of pre-trained models for specific constructs [22]. Furthermore, approaches like Knowledge Graph-enhanced Retrieval-Augmented Generation (KG-RAG) are emerging to specifically address the critical challenge of factual accuracy in LLM outputs [27]. Evaluating the effectiveness of these AI-generated items involves various methods, notably quantitative psychometric analyses focusing on properties like factor structure, reliability, difficulty, and discrimination [20, 22, 24], as well as qualitative expert and user reviews assessing content validity, relevance, fairness, and overall quality [23, 26, 25]. Despite the promise, a consistent theme across the studies is the necessity of expert review to address challenges such as potential inaccuracies [24, 27, 25], biases inherited from training data [20, 21, 26], and ensuring the over-

all quality and appropriateness of the generated items [22, 24, 26, 25]. The purpose of this discussion is therefore to interpret these findings in the specific context of generating effective soft skills MCQs, consider the broader implications, compare the observed methodologies, acknowledge limitations, and propose pertinent directions for future research.

Interpreting these findings for the generation of soft skills MCQs reveals both opportunities and significant challenges. The identified AI techniques—prompt engineering, fine-tuning, and potentially RAG—demonstrate clear pathways for generating assessment content using models like GPT-2, GPT-3, and ChatGPT [20, 22, 23, 24, 25]. Methods emphasizing construct-specificity through fine-tuning [22] or requiring precise, detailed instructions via prompt engineering [23, 25] appear particularly relevant for aligning generated MCQs with specific, often nuanced, soft skills competencies. Defining "effectiveness" for soft skills MCQs, based on the reviewed literature, likely necessitates a hybrid approach; quantitative psychometric properties like item difficulty and discrimination [24] are important for assessment utility, but qualitative expert judgment focusing on construct validity, contextual relevance, fairness, and potential bias [23, 26, 25] is arguably indispensable for ensuring the questions meaningfully address soft skills. The primary challenge lies in translating success from domains focused on cognitive abilities or stable traits [20, 22, 21, 23, 24, 26] to the dynamic, context-dependent, and often behavioral nature of soft skills like communication, teamwork, or problem-solving. It remains an open question how effectively the current methods can generate soft skills MCQs beyond surface-level knowledge. Techniques enhancing factual accuracy, such as KG-RAG [27], might prove valuable for creating reliable scenario-based questions common in soft skills training. Ultimately, the consistent finding across diverse studies [22, 24, 26, 25] that manual validation is essential underscores the interpretation that current AI technologies serve as powerful assistants rather than autonomous creators, particularly for complex and sensitive tasks like generating materials for soft skills development.

The implications of leveraging AI for soft skills MCQ generation are potentially far-reaching. Practically, the most significant implication, echoed across nearly all reviewed studies, is the potential for substantial efficiency gains, saving educators, trainers, and instructional designers considerable time and resources in developing practice materials and assessments [25, 26, 23, 22, 20]. This efficiency could facilitate the creation of larger, more diverse item banks [22], enabling personalized learning pathways or adaptive testing systems tailored to individual soft skills development needs. For the field of educational technology, these findings can guide the development of more sophisticated authoring tools, perhaps integrating user-friendly interfaces similar to the PIG tool concept [21] with robust back-ends for prompting, generation, and validation workflow support. Furthermore, this line of research has implications for AI development itself, highlighting the need for models with improved capabilities in handling nuance, context, ethical considerations, and bias mitigation – all critical aspects when dealing with soft skills. Regarding readiness for application, while studies show promise, such as achieving acceptable psychometric properties for medical MCQs [24], the strong emphasis on mandatory expert review [26, 25, 24, 22] suggests that AI-generated soft skills MCQs should perhaps be initially deployed in lower-stakes formative assessment or practice contexts, rather than high-stakes summative evaluations, pending further validation. The potential beneficiaries include learners gaining more practice opportunities, educators saving valuable time, and organizations obtaining more scalable tools for training and development.

Comparing the findings of the eight reviewed articles with the broader AIG landscape described by recent reviews [28, 29] highlights both continuity and rapid evolution. While [29] demonstrated that earlier AIG methods could produce high-quality MCQs comparable to traditional ones, even for assessing higher-order application of knowledge, the studies in this review predominantly showcase the capabilities of newer LLMs [20, 21, 22, 23, 24, 26, 25]. This focus aligns with the trend towards advanced Natural Language Processing (NLP) techniques noted [28]. Although these LLM-based approaches demonstrate significant potential for generating diverse items efficiently across various domains, the consistent emphasis across multiple articles [22, 24, 26, 25] on the necessity of rigorous expert validation due to concerns about accuracy [24, 27, 25] and bias [20, 21, 26] suggests that achieving the quality parity shown by [29] requires careful human-AI collaboration. The diverse evaluation methods employed in the reviewed articles—spanning psychometrics [20, 22, 24] and expert judgment [23, 26, 25]—reflect the varied evaluation landscape described by [28], reinforcing the importance of expert review also highlighted by [29] and underscoring the need for continued refinement of these powerful but still maturing AI techniques for operational use.

It is crucial to acknowledge the limitations present both within the reviewed studies and in this narrative review itself. A frequently reported limitation across the primary studies is the indispensable need for human expert review to ensure the quality, validity, fairness, and accuracy of the AI-generated items [22, 24, 26, 25]. Concerns regarding factual inaccuracies or "hallucinations" produced by LLMs were also noted [24, 27, 25], alongside the potential for inherent biases present in the models' training data influencing the generated content [20, 21, 26]. Specific studies reported issues like semantic redundancy or poor item quality [22], faced constraints due to small sample sizes or limited scope [24], or highlighted dependency on particular AI models or knowledge graphs [27]. Furthermore, many studies focused on domains other than soft skills or generated item formats other than MCQs (e.g., [20] focused on Likert-scale personality items). This review is also lim-

ited by its narrative nature and the small sample size of eight articles, which, despite the PRISMA methodology, may not capture the full breadth of research. Key questions pertinent to the research objective remain largely unanswered by this specific set of literature, including how to reliably generate MCQs that assess higher-order, context-dependent soft skills; how best to evaluate the effectiveness of these soft skills MCQs; and how to systematically mitigate biases in AI-generated content tailored for soft skills.

Based on the synthesis of findings and the identified limitations, several directions for future research emerge. There is a clear need for research directly comparing the efficacy of different AI techniques—prompt engineering, fine-tuning, RAG, or hybrid approaches—specifically tailored for the complexities of generating soft skills MCQs, potentially building upon the methods explored by [23, 22, 27]. Concurrently, robust evaluation frameworks must be developed for this specific context, integrating appropriate psychometric analyses with rigorous expert reviews that explicitly assess construct validity, contextual relevance, fairness, and potential biases, going beyond the general quality checks used in studies like [24, 26, 23]. Crucially, future empirical studies should move beyond evaluating item quality to investigating the actual impact of interacting with AI-generated soft skills MCQs on learners' skill development, ideally through longitudinal research designs. Addressing the persistent challenges of bias and inaccuracy is paramount; research should focus on developing and testing methods for proactively mitigating bias in soft skills content generation and enhancing the factual consistency of scenario-based questions, perhaps by refining KG-RAG techniques [27]. Finally, future work could explore the generation of diverse item formats suitable for soft skills, such as situational judgment test items (as suggested by [20]), investigate the integration of these tools into adaptive learning systems, and develop specific ethical guidelines for the responsible use of AI in soft skills assessment and development.

Overall, the reviewed studies suggests that LLMs offer considerable potential for streamlining MCQ generation, a potential that could, in principle, be extended to the domain of soft skills development. Techniques such as sophisticated prompt engineering, model fine-tuning, and retrieval-augmented generation represent key methodological avenues explored in related fields [23, 22, 27, 25]. However, the reviewed literature consistently underscores that realizing this potential effectively and responsibly hinges upon careful implementation and, critically, the integration of human expertise. Achieving desired levels of quality, accuracy, validity, and fairness currently necessitates a hybrid approach, with AI serving as a powerful assistant requiring human guidance, refinement, and validation [22, 24, 26, 25]. Significant challenges remain, particularly concerning the reliable generation and evaluation of content for nuanced soft skills and the mitigation of potential inaccuracies and biases. Addressing these challenges through focused future research on tailored techniques, robust evaluation methodologies, impact studies, and ethical frameworks will be essential to fully harness the capabilities of AI for soft skills MCQ generation. This review contributes by synthesizing the current state, highlighting both the promise and the necessary precautions based on recent studies investigating AI-driven AIG.

5. Conclusion

This narrative review examined the use of LLMs, particularly GPT-based architectures, for the automated generation of soft skills MCQs. The synthesis of eight recent studies identified core methodological trends, including prompt engineering, model fine-tuning, and RAG, primarily applied in domains such as cognitive assessment, medical testing, and language education. While these approaches demonstrate clear potential, their direct application to soft skills remains underexplored and methodologically challenging due to the inherently contextual and behavioral nature of such competencies.

Evidence from the reviewed literature suggests that LLMs can significantly increase the efficiency of AIG, particularly when integrated with structured prompting strategies or domain-specific fine-tuning. However, ensuring the quality, relevance, and fairness of generated items necessitates rigorous human oversight. Evaluation frameworks that combine psychometric analysis with expert review are essential, particularly when assessing content aimed at complex constructs like communication, teamwork, or ethical reasoning.

The primary implications concern both the scalability of AI-assisted assessment development and the limitations of current models. LLMs offer opportunities to expand item banks and support adaptive learning environments, but concerns about factual inaccuracies, semantic redundancy, and embedded biases highlight the importance of human-in-the-loop design. Current systems should therefore be deployed cautiously, preferably in formative or low-stakes contexts pending further validation.

This review is constrained by its limited sample size and the narrative synthesis methodology, which may not capture the full scope of relevant research. Additionally, most studies reviewed focus on domains other than soft skills or employ formats other than MCQs, limiting direct generalizability. Key gaps remain regarding the reliable generation of context-sensitive items, the empirical evaluation of learning outcomes, and the mitigation of AI-induced biases.

Future research should prioritize the development of specialized generation and evaluation frameworks for soft skills assessments. Comparative analyses of AI techniques, longitudinal impact studies, and bias mitigation strategies are particularly needed. As LLMs continue to evolve, their effective integration into soft skills development will depend on sustained interdisciplinary collaboration, combining advances in AI with domain-specific pedagogical and ethical expertise.

Ethics Statement

This study did not involve human participants or animals and therefore did not require ethical approval.

CRediT authorship contribution statement

Luis Cardena-Ley: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used ChatGPT to assist with language clarity and grammar correction. All final edits and intellectual content were determined and verified by the author.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G. W. Mitchell, L. B. Skinner, and B. J. White, "Essential soft skills for success in the twenty-first century workforce as perceived by business educators." *Delta Pi Epsilon Journal*, vol. 52, no. 1, 2010.

[2] A. F. Hendarman and U. Cantner, "Soft skills, hard skills, and individual innovativeness," *Eurasian Business Review*, vol. 8, pp. 139–169, 2018. doi: 10.1007/s40821-017-0076-6 .

[3] B. Schulz, "The importance of soft skills: Education beyond academic knowledge." *Journal of Language and Communication*, 2008.

[4] L. G. Malicay, "The integration of soft skills in professional education: Exploring the importance of communication, teamwork, and interpersonal skills in professional training and the methods used to incorporate them into educational programs," *International Journal of Advanced Research in Science, Communication and Technology*, vol. 3, no. 1, pp. 836–843, 2023. doi: 10.48175/IJARSCT-11967 .

[5] S. Maity and A. Deroy, "The future of learning in the age of generative ai: Automated question generation and assessment with large language models," *arXiv preprint arXiv:2410.09576*, 2024. doi: 10.48550/arXiv.2410.09576 .

[6] F. S. Mohammed and F. Ozdamli, "A systematic literature review of soft skills in information technology education," *Behavioral Sciences*, vol. 14, no. 10, p. 894, 2024. doi: 10.3390/bs14100894 .

[7] G. Fontaine, S. Cossette, M. A. Maheu-Cadotte, T. Mailhot, M. F. Deschênes, and G. Mathieu-Dupuis, "Effectiveness of adaptive e-learning environments on knowledge, competence, and behavior in health professionals and students: Protocol for a systematic review and meta-analysis," *JMIR Research Protocols*, vol. 6, no. 7, pp. 1–10, 2017. doi: 10.2196/resprot.8085 .

[8] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *bmj*, vol. 372, 2021. doi: 10.1136/bmj.n71 .

[9] R. Ferrari, "Writing narrative style literature reviews," *Medical writing*, vol. 24, no. 4, pp. 230–235, 2015. doi: 10.1179/2047480615Z.000000000329 .

[10] D. Torres-Salinas, R. Ruiz-Pérez, and E. Delgado-López-Cózar, "Google scholar como herramienta para la evaluación científica," *Profesional de la Información*, vol. 18, no. 5, pp. 501–510, 2009. doi: 10.3145/EPI.2009.SEP.03 .

[11] G. Halevi, H. Moed, and J. Bar-Ilan, "Suitability of google scholar as a source of scientific information and as a source of data for scientific evaluation—review of the literature," *Journal of informetrics*, vol. 11, no. 3, pp. 823–834, 2017. doi: 10.1016/j.joi.2017.06.005 .

[12] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *PLoS medicine*, vol. 6, no. 7, p. e1000097, 2009. doi: 10.1371/journal.pmed.1000097 .

[13] J. P. Higgins and S. Green, *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2011. doi: 10.1002/9780470712184 .

[14] A. Dobrescu, B. Nussbaumer-Streit, I. Klerings, G. Wagner, E. Persad, I. Sommer, H. Herkner, and G. Gartlehner, "Restricting evidence syntheses of interventions to english-language publications is a viable methodological shortcut for most medical topics: a systematic review," *Journal of Clinical Epidemiology*, vol. 137, pp. 209–217, 9 2021. doi: 10.1016/j.jclinepi.2021.04.012 . [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0895435621001347

[15] D. Gough, S. Oliver, and J. Thomas, *An introduction to systematic reviews*. SAGE, 2017. doi: 10.4135/9781036234942 .

[16] M. Callaham, R. L. Wears, and E. Weber, "Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals," *JAMA*, vol. 287, no. 21, pp. 2847–2850, 2002. doi: 10.1001/jama.287.21.2847.

[17] E. C. Martinez, P. E. G. Hasbun, V. P. S. Vargas, O. Y. García-González, M. D. F. Madera, D. E. R. Capistrán, T. C. Carmona, C. S. Cruz, and C. T. Hooper, "A comprehensive guide to conduct a systematic review and meta-analysis in medical research," *Medicine*, vol. 104, p. e41868, 8 2025. doi: 10.1097/MD.0000000000041868. [Online]. Available: https://journals.lww.com/10.1097/MD.0000000000041868

[18] N. R. Haddaway, P. Woodcock, B. Macura, and A. Collins, "Making literature reviews more reliable through application of lessons from systematic reviews," *Conservation Biology*, vol. 29, no. 6, pp. 1596–1605, 2015. doi: 10.1111/cobi.12541.

[19] A. V. Kulkarni, B. Aziz, I. Shams, and J. W. Busse, "Comparisons of citations in web of science, scopus, and google scholar for articles published in general medical journals," *JAMA*, vol. 302, no. 10, pp. 1092–1096, 2009. doi: 10.1001/jama.2009.1307.

[20] P. Lee, S. Fyffe, M. Son, Z. Jia, and Z. Yao, "A paradigm shift from "human writing" to "machine generation" in personality test development: an application of state-of-the-art natural language processing," *Journal of Business and Psychology*, vol. 38, pp. 163–190, 2023. doi: 10.1007/s10869-022-09864-6. [Online]. Available: https://doi.org/10.1007/s10869-022-09864-6

[21] F. M. Götz, R. Maertens, S. Loomba, and S. van der Linden, "Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development," *Psychological Methods*, vol. 29, pp. 494–518, 2024. doi: 10.1037/met0000540.

[22] B. E. Hommel, F. J. M. Wollang, V. Kotova, H. Zacher, and S. C. Schmukle, "Transformer-based deep neural language modeling for construct-specific automatic item generation," *Psychometrika*, vol. 87, pp. 749–772, 2022. doi: 10.1007/s11336-021-09823-9. [Online]. Available: https://doi.org/10.1007/s11336-021-09823-9

[23] U. Lee, H. Jung, Y. Jeon, Y. Sohn, W. Hwang, J. Moon, and H. Kim, "Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education," *Education and Information Technologies*, vol. 29, pp. 11 483–11 515, 2024. doi: 10.1007/s10639-023-12249-8. [Online]. Available: https://doi.org/10.1007/s10639-023-12249-8

[24] Y. S. Kıyak, Özlem Coşkun, I. İrem Budakoğlu, and C. Uluoğlu, "Chatgpt for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam," *European Journal of Clinical Pharmacology*, vol. 80, pp. 729–735, 2024. doi: 10.1007/s00228-024-03649-x.

[25] I. R. Indran, P. Paranthaman, N. Gupta, and N. Mustafa, "Twelve tips to leverage ai for efficient and effective medical question generation: A guide for educators using chat gpt," *Medical Teacher*, vol. 46, pp. 1021–1026, 2024. doi: 10.1080/0142159X.2023.2294703.

[26] Y. Attali, A. Runge, G. T. Laflair, K. Yancey, S. Goodwin, Y. Park, and A. A. V. Davier, "The interactive reading task: Transformer-based automatic item generation," *Frontiers in Artificial Intelligence*, vol. 5, 7 2022. doi: 10.3389/frai.2022.903077.

[27] K. Soman, P. W. Rose, J. H. Morris, R. E. Akbas, B. Smith, B. Peetoom, C. Villouta-Reyes, G. Cerono, Y. Shi, A. Rizk-Jackson, S. Israni, C. A. Nelson, S. Huang, and S. E. Baranzini, "Biomedical knowledge graph-optimized prompt generation for large language models," *Bioinformatics*, vol. 40, 11 2024. doi: 10.1093/bioinformatics/btae560. [Online]. Available: http://arxiv.org/abs/2311.17330

[28] R. Circi, J. Hicks, and E. Sikali, "Automatic item generation: foundations and machine learning-based approaches for assessments," 2023. doi: 10.3389/feduc.2023.858273.

[29] D. Pugh, A. D. Champlain, M. Gierl, H. Lai, and C. Touchie, "Can automated item generation be used to develop high quality mcqs that assess application of knowledge?" *Research and Practice in Technology Enhanced Learning*, vol. 15, 2020. doi: 10.1186/s41039-020-00134-8.