



BEYOND FLUENCY: COGNITIVE CONTROL AND PEDAGOGICAL QUALITY IN LLM-BASED AUTOMATIC QUESTION GENERATION — A SYSTEMATIC REVIEW OF BLOOM-ALIGNED ASSESSMENT PRACTICES (2020–2025)

Firmansyah Ibrahim	Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia <i>and</i> Department of Information Systems, Universitas Islam Negeri Alauddin Makassar, Makassar, Indonesia	firmansyah.ibrahim.2505349@students.um.ac.id
Didik Dwi Prasetya*	Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia	didikdwi@um.ac.id
Triyanna Widiyaningtyas	Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia	triyannaw.ft@um.ac.id

* Corresponding author

ABSTRACT

Aim/Purpose	This systematic review synthesises 69 primary studies (2020–2025) to map how contemporary automatic question generation (AQG) research operationalises cognitive control, quality validation, and language inclusivity, and to identify where the field still needs to make progress.
-------------	---

Accepting Editor Digna Sayco Evale | Received: April 4, 2026 | Revised: June 3, August 25, 2026 |
Accepted: September 16, 2026.

Cite as: Ibrahim, F., Prasetya, D. D., & Widiyaningtyas, T. (2026). Beyond fluency: Cognitive control and pedagogical quality in LLM-based automatic question generation — A systematic review of Bloom-aligned assessment practices (2020–2025). *Journal of Information Technology Education: Research*, 25, Article 40.
<https://doi.org/10.28945/5887>

(CC BY-NC 4.0) This article is licensed to you under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/). When you copy and redistribute this paper in full or in part, you need to provide proper attribution to it to ensure that others can later locate this work (and to ensure that others do not accuse you of plagiarism). You may (and we encourage you to) adapt, remix, transform, and build upon the material for any non-commercial purposes. This license does not permit you to use this material for commercial purposes.

Background	While large language models (LLMs) have resolved the fluency challenge in automatic question generation (AQG), they have not resolved the harder pedagogical challenge. The field must now shift from “fluent question generation” to “pedagogically valid question generation,” creating assessments that strictly align with cognitive frameworks such as Bloom’s Taxonomy and possess verifiable docimological quality.
Methodology	A systematic literature review (SLR) was conducted across Scopus, IEEE Xplore, and SpringerLink, filtering 775 initial records to 69 primary studies (2020–2025) using PRISMA-guided inclusion/exclusion and quality assessment criteria.
Contribution	Unlike prior AQG reviews that address technical performance or single pedagogical dimensions in isolation, this review is the first to systematically integrate three concerns: (1) Bloom-aligned cognitive control, (2) pedagogical validation beyond automated metrics, and (3) language inclusivity. By synthesising these pillars, the study provides a taxonomy of control methods, establishes expert evaluation as the validation reference standard, highlights the low-resource language gap (e.g., Indonesian), and proposes a Pedagogical Alignment Framework.
Findings	The review reveals a clear architectural divide between prompt-based flexibility and fine-tuning precision. The data show that 88.4% of the studies addressed cognitive control and that all 69 studies (100%) applied some form of quality validation beyond n-gram metrics. However, only 24.6% did so with full docimological rigour. In addition, 87.0% of the studies focused exclusively on English-language contexts. The review also establishes that automated metrics such as BLEU and ROUGE are inadequate for measuring pedagogical quality.
Recommendations for Practitioners	To ensure docimological quality, practitioners must mandate subject matter expert (SME) validation. Furthermore, institutional deployment should strategically align the AI architecture with the learning goal: choosing parameter-efficient fine-tuning (PEFT) for strict cognitive compliance (e.g., high-stakes exams) or LLM prompting for flexible, exploratory tasks.
Recommendations for Researchers	Researchers and developers are encouraged to focus on creating datasets for non-English and low-resource languages to address equity gaps in educational technology.
Impact on Society	By addressing the cognitive and validation challenges in AQG, this review promotes the development of more reliable, pedagogically sound, and equitable AI-supported assessment tools for diverse learner populations.
Future Research	Future research should focus on empirically validating the proposed Pedagogical Alignment Framework, developing automated assessors to relieve the validation bottleneck, and expanding AQG applications in culturally contextualised and low-resource languages.
Keywords	automatic question generation, large language models, Bloom’s Taxonomy, quality validation, systematic literature review

INTRODUCTION

The main obstacle in automatic question generation (AQG) today is no longer linguistic fluency. It is the generation of assessments that satisfy the requirements of pedagogical validity and of specific educational needs (Bulathwela et al., 2023; Navas et al., 2025). L. Wang et al. (2025) and Winata et al. (2025) describe the development of high-quality assessment items as a demanding, time-consuming, and inherently creative pedagogical task. In practice, assessment creation represents a significant and recurring burden for educators working across learning management systems, adaptive tutoring systems, and classroom environments, a burden that AQG systems are well-positioned to reduce, provided that the questions generated are not only fluent but pedagogically sound (Arunnehr et al., 2025; Van Campenhout et al., 2021).

The field has progressed from inflexible rule-based systems to sophisticated deep learning frameworks, moving from the early application of transfer learning (Liao & Koh, 2020) to more recent Transformer-based models and large language models (LLMs) (Agarwal et al., 2022; Suhartono et al., 2024). These advances largely resolved the early challenge of fluency and made it far easier to generate multiple-choice questions (MCQs) with plausible distractors (Kauchak et al., 2024; Yin & Li, 2025). They also redirected the research agenda. The goal is no longer grammatical correctness alone, but rather questions that carry genuine educational value, and reaching that goal entails overcoming three interrelated barriers: cognitive demand (Elkins et al., 2023; Zichao Wang et al., 2022), pedagogical validation, and linguistic inclusivity for low-resource languages.

Cognitive control is the first of these challenges. Standard AQG systems produce simple, factual questions (Scaria et al., 2024). In practice, however, assessment must span the full cognitive domain, particularly higher-order thinking skills. For this reason, Bloom’s Taxonomy has become the default educational framework for guiding and constraining LLM outputs (Jadhav et al., 2025; Timosh et al., 2025). This systematic literature review (SLR) found that most primary studies apply this control on two fronts: (1) prompt engineering applied to general-purpose LLMs, which is the more flexible option and requires no additional training (Bahri & Oueslati, 2025; Lee et al., 2024), and (2) fine-tuning of encoder–decoder models such as T5, which offers tighter control because the cognitive target can be tuned through the model parameters (Atmajaya et al., 2025; Suhartono et al., 2024).

Quality validation is the second challenge and no less important. The task is to establish the validity of questions generated by LLMs, at least with respect to their cognitive demand, factual relevance, and educational appropriateness. Our review indicates a consensus that BLEU, ROUGE, and similar fully automated metrics do not capture pedagogical quality (Elkins et al., 2023; Grévisse, 2024). This explains why evaluation by subject matter experts (SMEs) is considered the reference standard (Van Campenhout et al., 2021; Wróblewska et al., 2025). Such evaluations focus on assessment validity (Grévisse et al., 2024) and are sometimes supplemented by psychometric models such as Item Response Theory (IRT) to predict difficulty levels (J. Feng & Mostow, 2021; Tomikawa & Uto, 2024).

Despite rapid technological advancements, the benefits of controlled and validated AQG remain unevenly distributed because of severe inclusivity gaps. Research to date has centred on English and on high-stakes, resource-intensive professional fields such as medicine and computer science (Maity et al., 2024; Mulla & Gharpure, 2023). This linguistic gap is not merely a matter of coverage; it also extends to the control and validation challenges identified above: low-resource languages lack both the annotated corpora needed for Bloom-level fine-tuning and the subject-matter expert pools needed for pedagogical validation. Only a few pioneering studies have investigated this intersection by implementing cognitive control and quality validation for AQG in Indonesian (Atmajaya et al., 2025; Fuadi & Wibawa, 2022; Suhartono et al., 2024; Vincentio & Suhartono, 2022) and other non-English languages such as Thai (Wiwatbutsiri et al., 2022) and Vietnamese (D. H. Nguyen et al., 2025). This review synthesises 69 primary studies (2020–2025) to map the current state of the art in AQG. Specifically, the main contributions of this review are to: (1) map the landscape of low-resource language research with a focus on the Indonesian gap, (2) classify advanced cognitive control methods aligned

with Bloom’s Taxonomy, (3) examine pedagogical quality validation practices beyond automated metrics, (4) identify dominant generative models and application domains, and (5) synthesise these findings into a comprehensive Pedagogical Alignment Framework.

RELATED WORK

To ensure conceptual clarity, several foundational terms are defined here. Cognitive control refers to the mechanism for steering the AI generation process to produce questions at targeted levels of cognitive complexity, as defined by Bloom’s Taxonomy. Assessment validity denotes the extent to which an assessment instrument accurately measures the specific learning constructs it is intended to evaluate (Messick, 1995). Docimological quality, a central concept in the science of educational examination, is defined here as the overall quality of an assessment item in terms of its validity, fairness, clarity, distractor plausibility, and appropriateness for the intended cognitive level. Finally, pedagogical alignment refers to the essential congruence among the generated assessment items, the learning objectives, and the instructional framework.

Before turning to recent advances in generative AI, it is important to situate AQQ within its foundational pedagogical and technical frameworks. The cognitive goals of contemporary AQQ systems can be mapped onto the original taxonomy by Bloom et al. (1956) and its extension by Anderson and Krathwohl (2001). The current focus on the pedagogical value of LLM outputs aligns closely with the early concerns of educational measurement, specifically construct validity, which requires that test items be representative of the learning goals (Messick, 1995). Early efforts in computer science to automate the assessment process, primarily from heuristic and rule-based perspectives, constitute the technical point of departure before the Transformer era (Kurdi et al., 2020). Although these early systems established the foundations of the field, their limited semantic flexibility highlighted the need for deep learning models to achieve pedagogical alignment.

The first wave of AQQ research after the advent of the Transformer architecture combined linguistic feature extraction with early neural approaches, thereby confirming the efficacy of deep learning in producing MCQs and plausible distractors (Agarwal et al., 2022; Suhartono et al., 2024). Attention then turned to larger, more scalable pre-trained language models (PLMs) and prompt learning, which accelerated the generation process (Bulathwela et al., 2023; Wu et al., 2023). Subsequent success, however, created a problem of a different order: control. Recent studies have highlighted the difficulty of shaping an LLM’s output to meet specific educational objectives (L. Wang et al., 2025). Figure 1 maps the keyword co-occurrence structure of the retrieved literature (map parameters are reported in the Methodology section). The dense core of the field comprises generation-oriented terms such as semantics, natural language generation, and language models, whereas terms denoting pedagogical control and validation occupy peripheral positions. This imbalance is the point of departure for the present review.

The first pedagogical integrations began to address this control issue. Bloom’s Taxonomy became the dominant framework for steering LLM output, as it was the most widely used scheme for eliciting questions across the cognitive levels, from C1 (remembering) to C6 (creating) (Jadhav et al., 2025; Scaria et al., 2024; Timosh et al., 2025). Quality validation emerged as a second-order challenge. Once controlled generation became feasible, researchers turned to the question of whether the resulting items were usable in practice. Early evaluations found that LLM-generated questions were consistently fluent, but fluency proved to be a poor proxy for pedagogical soundness: items that read well frequently failed to address the intended educational construct. A consensus, therefore, emerged that conventional automated measures are insufficient for judging question quality (Grévisse, 2024), and subsequent work shifted towards purpose-built evaluation procedures centred on assessment validity, a property that only domain specialists can reliably establish (Grévisse et al., 2024).

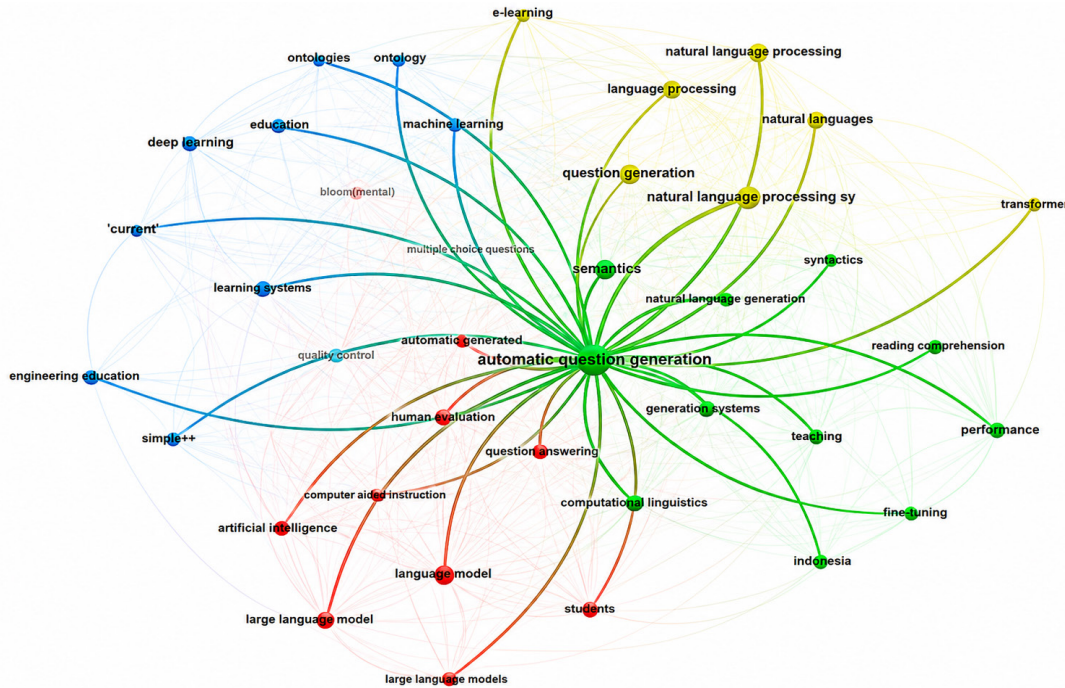


Figure 1. Keyword co-occurrence map of the AQG literature, centred on the focal term “automatic question generation” (VOSviewer)

To bring together the scattered analyses on technical architectures, assessment constraints, and cognitive frameworks, we synthesise the core literature in Table 1. Whereas earlier analyses typically examine studies individually, this comparison groups the most advanced AQG methodologies by generation mechanism (e.g., fine-tuning, prompting), by reliance on n-grams versus docimological assessment, and by incorporation of Bloom’s Taxonomy. The comparison underlines the need for an integrated pedagogical and technical framework.

Table 1. Critical synthesis of state-of-the-art AQG approaches

AQG generation approach	Cognitive control strategy	Quality validation method	Representative studies	Critical synthesis and limitations
Fine-tuning (pre-trained language models)	Implicit or data-driven (no explicit taxonomy)	Automated surface-level metrics (BLEU, ROUGE, METEOR)	Agarwal et al. (2022); Lamsiyah et al. (2024); Suhartono et al. (2024)	Produces high syntactic fluency and grammatical correctness. However, n-gram metrics fail to capture construct validity, so the pedagogical utility of the questions remains uncertain.
Prompt-based LLMs (zero-shot and few-shot)	Heuristic constraints (prompt engineering)	Mixed (automated and limited human review)	Bulathwela et al. (2023); L. Wang et al. (2025); Wu et al. (2023)	Accelerates generation and offers broad domain coverage. It suffers, however, from significant control problems: models frequently hallucinate or fail to adhere strictly to the requested cognitive level without structured fine-tuning.

AQG generation approach	Cognitive control strategy	Quality validation method	Representative studies	Critical synthesis and limitations
Taxonomy-aligned and adaptive systems	Explicit mapping to Bloom's Taxonomy	Human expert and docimological validation; IRT	Grévisse et al. (2024); Jadhav et al. (2025); Scaria et al. (2024); Timosh et al. (2025)	Achieves strong pedagogical alignment and verifies educational value. However, human validation is resource-intensive, and such control mechanisms remain underexplored for low-resource languages.

Beyond architectural advancements, the literature highlights a core shift in how AQG is understood within educational systems. To bring the current state of the art together, we group the literature into three dimensions.

1. **AQG as a driver for active learning.** The literature indicates that AQG has evolved from a tool for efficiency into a tool for engagement. Van Campenhout et al. (2021) demonstrated the “doer effect,” in which embedding automatically generated questions in reading materials converted passive reading into active learning: students showed higher engagement and greater persistence than students who read the same material without embedded questions. Arunnehr et al. (2025) extended this line of work by integrating AQG into intelligent learning assistants that generate personalised study plans and support self-regulated learning.
2. **Effects of question quality on learning outcomes.** L. Wang et al. (2025) show that what matters for learning outcomes is the pedagogical quality of questions, not their readability alone: well-constructed items support comprehension and higher-order thinking, whereas poorly constructed ones create confusion and depress performance. Steuer et al. (2021) likewise argue that grammatical well-formedness is not sufficient for validity. Whether an item supports the self-assessment that is the goal of the learning activity depends on its educational substance rather than its surface form.
3. **Towards a unified pedagogical and technical framework.** The literature points to an emergent framework in which technical architecture and pedagogical control must be aligned. Susanti et al. (2020) were among the first to integrate AQG into computerised adaptive testing (CAT). They showed that technical accuracy (fine-tuning) enables a system to adjust dynamically to a learner’s ability level. This positions generative AI as an adaptive system that modulates the complexity of assessment tasks in line with the learner’s cognitive progression.

Taken together, these three dimensions – the motivational role of AQG in active learning, the link between question quality and learning outcomes, and the technical imperative of pedagogical alignment – establish the theoretical foundation of this review. A study that maps both the control mechanisms and the validation practices used throughout the current literature is therefore not merely timely but also theoretically necessary.

Three previous reviews provide useful points of comparison. First, Kurdi et al. (2020) provided an extensive survey of AQG methods but predated the LLM era; thus, they did not address Bloom-aligned cognitive control or pedagogical validation. Second, Pan et al. (2019) focused on the underlying architecture of question generation strictly as a natural language processing (NLP) task, without considering educational validity frameworks. Finally, while Elkins et al. (2023) assessed LLM-generated questions and their use in teaching, they addressed neither cognitive control mechanisms nor

low-resource languages. The present review is the first to examine the integration of cognitive control, pedagogical validation, and linguistic inclusivity within the post-Transformer (2020–2025) period.

The primary studies reviewed here confirm that this research focus has matured. Although control has been identified as a significant challenge (L. Wang et al., 2025) and validation as the corresponding requirement (Grévisse et al., 2024), no systematic literature review has yet synthesised both pillars. This SLR fills that gap by mapping both pillars across the 69 selected primary studies.

Figures 2 and 3 present the corresponding maps for the two remaining search themes. In Figure 2, “bloom” appears as a low-frequency node linked to the AQG core, indicating that cognitive control is present in the literature but not yet a dominant term. In Figure 3, “human evaluation” appears as a distinct node linked directly to the AQG core, anticipating the validation practices analysed under RQ3. The maps are read here descriptively, as a characterisation of the retrieved literature; the quantitative findings reported in the Results section are derived from the manual coding of the 69 primary studies.

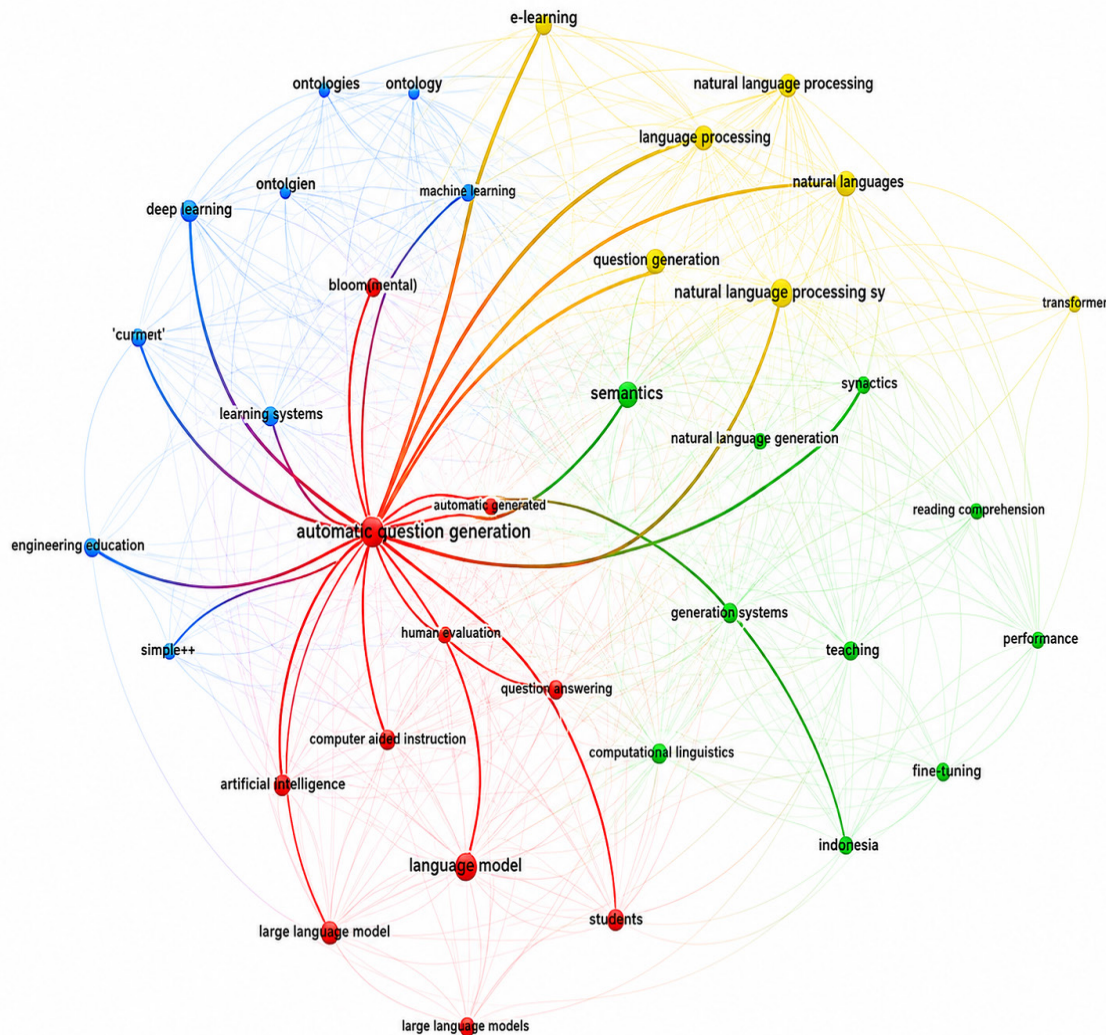


Figure 2. Keyword co-occurrence map for the cognitive-control search theme, showing the position of “bloom” relative to the AQG core (VOSviewer)

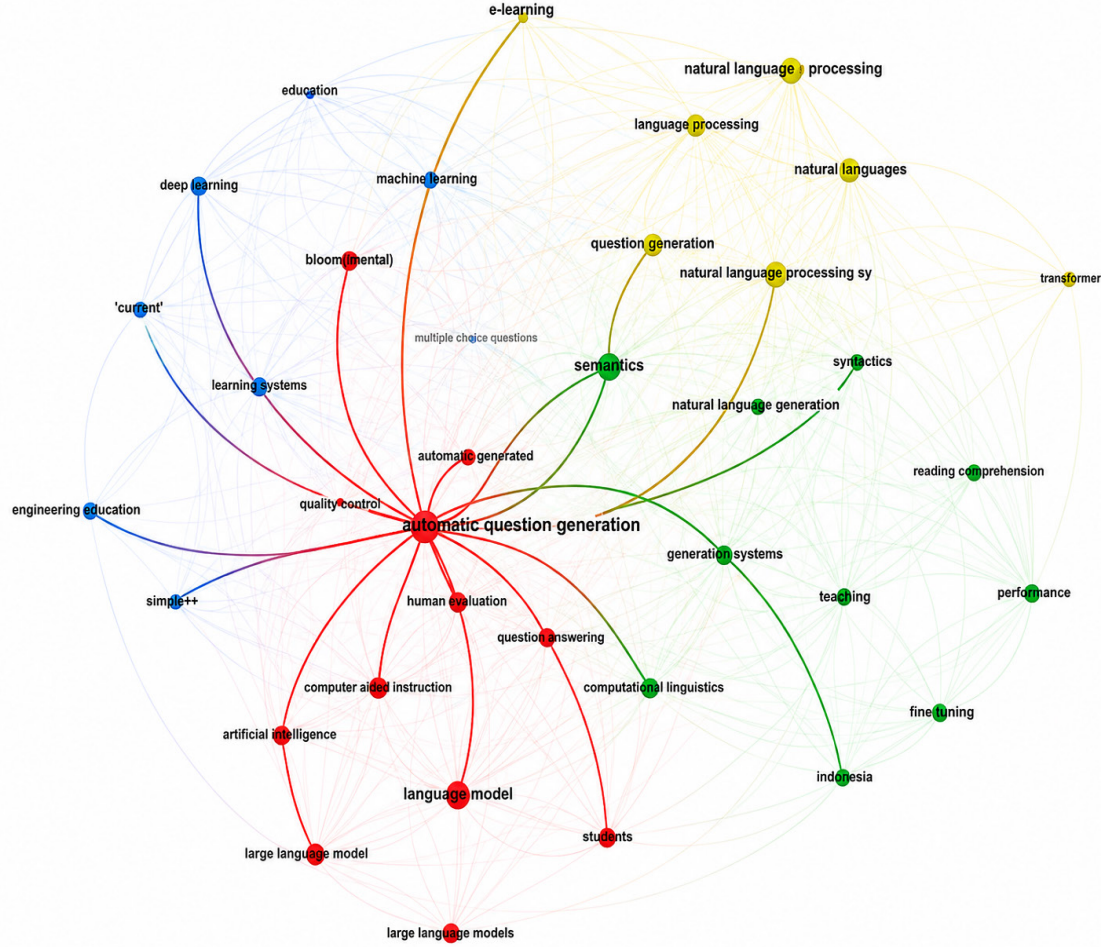


Figure 3. Keyword co-occurrence map for the validation search theme, showing the position of “human evaluation” relative to the AQQ core (VOSviewer)

METHODOLOGY

This SLR was conducted following standard guidelines and protocols, encompassing three distinct phases: (1) planning, (2) conducting, and (3) reporting.

PLANNING

The planning phase began by defining the scope of the SLR. That scope was deliberately narrow: rather than reviewing AQQ in general, we focused on the intersection of three concerns, namely cognitive control through frameworks such as Bloom’s Taxonomy, quality validation beyond standard automated metrics, and recent implementations based on large language models and Transformer architectures. On this basis, the research questions (RQs) were defined as follows:

- (RQ1) What is the landscape of AQQ research for low-resource languages, with a specific focus on Indonesian as a case study?
- (RQ2) What are the current advanced methods used for cognitive control, particularly using Bloom’s Taxonomy?
- (RQ3) What are the quality validation methods used to assess pedagogical quality?
- (RQ4) What are the main application domains and the dominant LLM models employed?

Data Sources and Search Strategy

Our focus on three pillars – language, control, and validation – required an approach that differed from that of conventional SLRs. We used a multi-stage search strategy to obtain broad coverage. We applied three distinct search queries across Scopus, IEEE Xplore, and SpringerLink to identify primary studies published between 2020 and 2025 (Figure 4).

The keyword co-occurrence maps shown in Figures 1 to 3 were generated in VOSviewer using records returned by these queries. Author and index keywords were merged using a thesaurus file, and low-frequency keywords were excluded so that only terms recurring across multiple records were retained. Node size is proportional to occurrence frequency, link thickness to co-occurrence strength, and colour denotes the cluster assigned by the association-strength algorithm; the interpretation offered in the text does not depend on colour. The maps serve a descriptive purpose only, namely to characterise the structure of the retrieved literature. All quantitative findings reported in the Results were obtained through manual coding of the 69 primary studies, not from the maps.

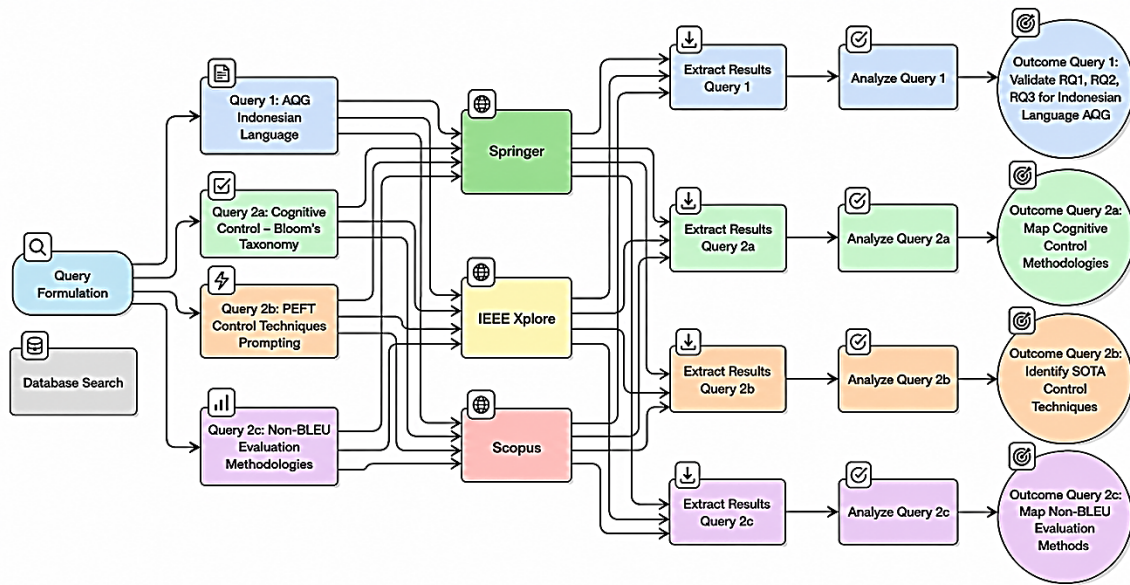


Figure 4. Data sources and search strategies

The search protocol was designed to retrieve studies addressing the three pillars of control, validation, and language. Each query was restricted to the title, abstract, and keyword fields (TITLE-ABS-KEY in Scopus) and required the mandatory combination of core terms (“automatic question generation” OR “AQG” OR “question generation”). This core was then combined with four groups of terms. To retrieve contemporary (post-Transformer) implementations, it was combined with “LLM” OR “prompt” OR “fine-tuning” OR “PEFT”. Cognitive control was targeted using “Bloom” OR “cognitive level” OR “difficulty,” reflecting a shift in research from fluency to pedagogical adherence. Quality validation was targeted through “human evaluation” OR “faithfulness”, which distinguishes studies using expert assessment as the reference standard from those relying on conventional automated metrics. Finally, the linguistic gap was addressed by using “Indonesian” or “Bahasa Indonesia” to retrieve relevant local studies.

To support reproducibility, the database searches were conducted from November 5 to 7, 2025. Publication years recorded in this review refer to the year of first online publication, as several included studies were in advance online form at the time of screening. Database-specific syntax was written for each source. The full search strings for Scopus, IEEE Xplore, and SpringerLink, organised by the main research themes, are presented in Table 2.

Table 2. Search strings by database

Search theme	SpringerLink syntax	IEEE Xplore syntax	Scopus syntax
1. AQG and Indonesian language	((“automatic question generation” OR “question generation” OR “AQG”) AND (“Indonesian” OR “Bahasa Indonesia”))	((“automatic question generation” OR “question generation” OR “AQG”) AND (“Indonesian” OR “Bahasa Indonesia”))	TITLE-ABS-KEY ((“automatic question generation” OR “question generation” OR “AQG”) AND (“Indonesian” OR “Bahasa Indonesia”))
2a. Cognitive control and Bloom’s Taxonomy	((“automatic question generation” OR “AQG”) AND (“controllable” OR “control” OR “controlling”) AND (“Bloom’s Taxonomy” OR “cognitive level” OR “difficulty level” OR “educational”))	((“automatic question generation” OR “AQG”) AND (“controllable” OR “control” OR “controlling”) AND (“Bloom’s Taxonomy” OR “cognitive level” OR “difficulty level” OR “educational”))	TITLE-ABS-KEY ((“automatic question generation” OR “AQG”) AND (“controllable” OR “control” OR “controlling”) AND (“Bloom’s Taxonomy” OR “cognitive level” OR “difficulty level” OR “educational”))
2b. Control techniques (PEFT and prompting)	((“automatic question generation” OR “controllable generation” OR “AQG”) AND (“instruction tuning” OR “prompting” OR “PEFT” OR “LORA” OR “prefix-tuning” OR “adapters”))	((“automatic question generation” OR “controllable generation” OR “AQG”) AND (“instruction tuning” OR “prompting” OR “PEFT” OR “LORA” OR “prefix-tuning” OR “adapters”))	TITLE-ABS-KEY ((“automatic question generation” OR “controllable generation” OR “AQG”) AND (“instruction tuning” OR “prompting” OR “PEFT” OR “LORA” OR “prefix-tuning” OR “adapters”))
2c. Evaluation methodology (non-BLEU)	((“automatic question generation” OR “AQG”) AND (“evaluation” OR “metric” OR “protocol”) AND (“human evaluation” OR “expert” OR “faithfulness” OR “relevance” OR “taxonomy” OR “compliance” OR “cognitive”))	((“automatic question generation” OR “AQG”) AND (“evaluation” OR “metric” OR “protocol”) AND (“human evaluation” OR “expert” OR “faithfulness” OR “relevance” OR “taxonomy” OR “compliance” OR “cognitive”))	TITLE-ABS-KEY ((“automatic question generation” OR “AQG”) AND (“evaluation” OR “metric” OR “protocol”) AND (“human evaluation” OR “expert” OR “faithfulness” OR “relevance” OR “taxonomy” OR “compliance” OR “cognitive”))

CONDUCTING

The conducting phase began by applying the search queries to the selected databases. This yielded 775 initial records (SpringerLink: $n = 611$; Scopus: $n = 143$; IEEE Xplore: $n = 21$). The number of records returned by SpringerLink was disproportionately high compared with those returned by the indexing database Scopus. This discrepancy arose because the syntax used in Scopus was limited to the TITLE-ABS-KEY (title, abstract, and keywords) fields, which ensures high topical relevance, whereas the search interface in SpringerLink defaults to a broader full-text search. The SpringerLink results, therefore, included a large volume of loosely related documents that mentioned the search terms only in passing.

The next step was to filter these results systematically to identify the relevant primary studies; the complete screening flow is summarised in the PRISMA flow diagram (Figure 5). We first screened titles and abstracts against the established inclusion and exclusion criteria. A single reviewer conducted title/abstract screening and full-text eligibility assessment due to resource constraints. To minimise potential selection bias, all screening decisions followed predefined inclusion and exclusion criteria, were guided by the PRISMA protocol, and were supported by a structured quality assessment rubric applied consistently across all studies. Nevertheless, because no independent second reviewer

was involved, inter-rater agreement statistics could not be calculated. The complete criteria utilised to filter the documents are detailed in Table 3.

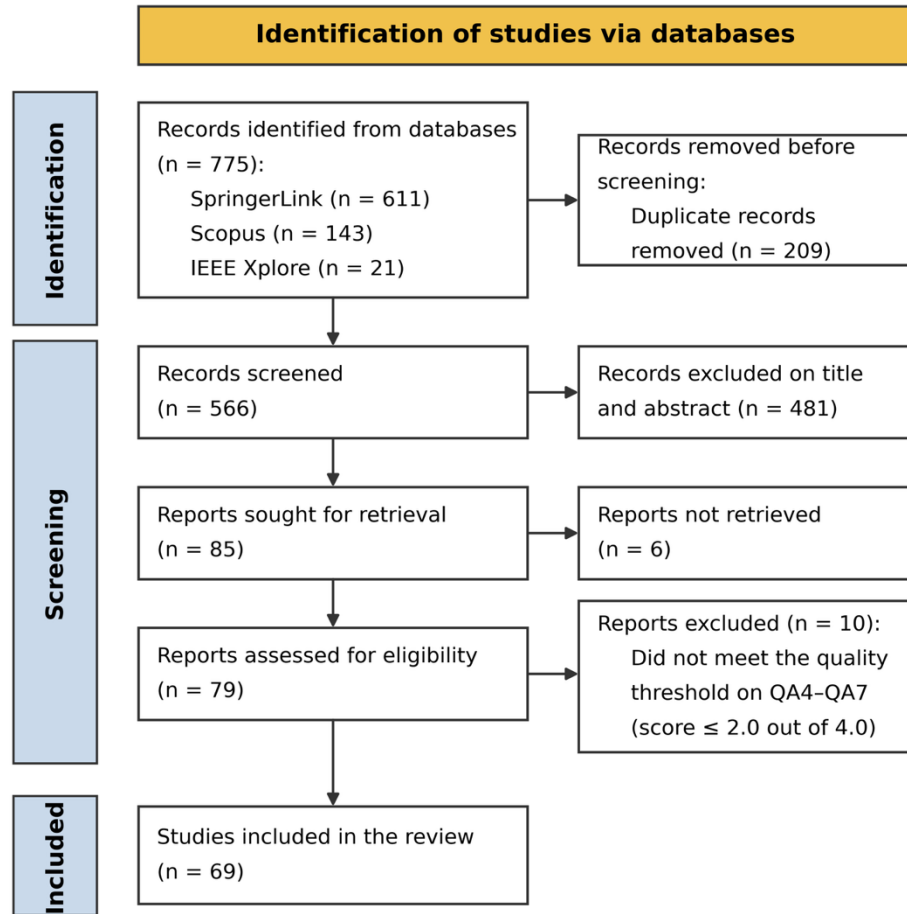


Figure 5. PRISMA flow diagram of the study selection process

Table 3. SLR inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Focuses explicitly on automatic question generation (AQG).	Does not focus primarily on AQG (e.g., general text summarisation or generic question answering).
Applies modern pre-trained language models, Transformer architectures, or large language models (LLMs).	Focuses solely on older, rule-based, or non-Transformer NLP models.
Explicitly discusses or applies cognitive control methods (e.g., Bloom's Taxonomy, difficulty levels, or pedagogical alignment).	Lacks explicit control mechanisms for cognitive demand or difficulty levels.

Inclusion criteria	Exclusion criteria
Applies quality validation methods beyond standard automated metrics (e.g., human evaluation, SME validation, or docimological quality assessment).	Relies exclusively on automated n-gram metrics (e.g., BLEU, ROUGE) without any pedagogical or manual validation.
Written in English or Indonesian.	Written in languages other than English or Indonesian.
Published between 2020 and 2025 (the post-Transformer and LLM era).	Published before 2020.
Primary empirical study (journal article or conference paper) reporting a dataset or corpus and experimental results.	Secondary study (e.g., literature review, survey) or theoretical paper lacking experimental data.

After removing duplicates and irrelevant documents, we obtained 79 potentially relevant studies. We then examined these 79 studies in depth by reading their full texts and assessing them against a set of quality assessment (QA) criteria. These criteria were designed to ensure that each study genuinely addressed cognitive control (RQ2) and quality validation (RQ3), thereby maintaining a consistent methodological standard.

To assess the quality and relevance of these studies, we applied a QA protocol consisting of seven criteria:

- (QA1) The study reports primary empirical research.
- (QA2) The study focuses on automatic question generation.
- (QA3) The study employs a Transformer-based or LLM architecture.
- (QA4) The study explicitly discusses cognitive control methods, such as Bloom’s Taxonomy.
- (QA5) The study reports quality validation methods, such as expert evaluation or rubrics.
- (QA6) The study describes its corpus or dataset.
- (QA7) The study presents its experimental findings clearly.

The final 69 studies were selected through a two-tiered quality assessment. First, every candidate study was scored against the QA criteria (1 = yes, 0.5 = partially, 0 = no). To be considered further, a study had to score full marks on the mandatory criteria (QA1–QA3); this served as the first filter. It then had to pass a quality threshold of more than 2.0 out of 4.0 on the remaining criteria (QA4–QA7). This procedure removed 10 studies. The most common reason for exclusion was that a study addressed AQG but lacked the required methodological quality (score ≤ 2.0). The process yielded a final corpus of 69 studies for analysis. For full methodological transparency, the complete QA rubric, the operational definitions for each score level, the rationale for the inclusion threshold, and the detailed QA scores for all 69 included studies are provided in the Appendix.

REPORTING

After the primary study selection process was completed, the data extraction phase was conducted. The data extraction form (Table 4) was designed to collect relevant information and address each research question systematically. To process the extracted information from the 69 included studies, a hybrid data synthesis approach was employed. First, a descriptive statistical synthesis was used to quantify trends in publication frequencies, model architectures, and validation types. Second, a narrative and thematic synthesis was conducted to categorise and interpret the qualitative aspects of the studies, specifically grouping the findings into three core research themes: cognitive control methods

(identifying specific prompt-engineering and fine-tuning taxonomy alignments); validation strategies (comparing automated metrics with docimological quality); and linguistic focus. This combined approach supports both a quantitative mapping of the current research landscape and a qualitative understanding of the pedagogical alignment frameworks employed. The analysis and synthesis of the findings from this extraction process are presented in detail in the Results section.

Table 4. Description of the categories analysed

Category	Description
Publication details	General information about the selected studies, including the year of publication and the publisher.
Application domain	The specific subject area or educational context where the AQG is applied (e.g., medicine, computer science, law, or general education).
Model architecture	The type of generative AI model used, distinguishing between encoder–decoder Transformer models (e.g., T5, BERT) and large-scale LLMs (e.g., GPT-series, Llama).
Control methodology	The specific technical approach used to achieve controlled generation, categorised into fine-tuning strategies (e.g., full fine-tuning, parameter-efficient fine-tuning, LoRA) or prompt engineering strategies (e.g., zero-shot, few-shot, chain-of-thought, retrieval-augmented generation).
Cognitive framework	The pedagogical standard or framework targeted by the control methodology (e.g., Bloom’s Taxonomy levels, Item Response Theory, or general difficulty levels).
Input context	The characteristics of the input source provided to the model for question generation (e.g., plain text, textbook documents, structured format).
Validation strategy	The evaluation metrics and protocols employed to assess the generated questions, specifically identifying the presence of manual pedagogical metrics (e.g., expert judgement, docimological quality, factual accuracy) versus automated metrics (e.g., BLEU, ROUGE).

RESULTS

This SLR includes 69 primary studies that passed the QA stage described in the previous section. Figure 6 presents a summary of the results for each QA item across these 69 primary studies.

The mandatory criteria (QA1–QA3) were met by all included studies. Scores on the quality criteria were also high: QA4 (cognitive control) was met (score ≥ 0.5) by 61 studies (88.4%), of which 10 (14.5%) met it fully, and QA5 (quality validation) was met by all 69 studies (100%), of which 17 (24.6%) met it fully and 52 (75.4%) only partially. These results confirm the relevance of the 69 primary studies to the two pillars of this review: cognitive control and quality validation. The lowest score, although still high, was recorded for QA4, indicating that some AQG studies still focus on general generation without explicit cognitive control (Jadhav et al., 2025).

The primary studies span 2020 to April 2025 and show a marked acceleration in research activity after 2022. This growth, shown in Figure 7, followed the release of major LLMs, such as the GPT series, and was accompanied by a shift towards more advanced control and evaluation methods (Lamsiyah et al., 2024; Wu et al., 2023). By publication venue, conference proceedings dominated (44 studies, 63.8%), followed by journal articles (23 studies, 33.3%) and preprints (2 studies, 2.9%).

Furthermore, as shown in Figure 6, the high score for QA4 (addressing cognitive control) at 88.4% confirms that control is the primary focus of current AQG research. Our analysis of the 69 primary studies categorised control methods along two dimensions: (1) control frameworks (e.g., Bloom’s Taxonomy versus difficulty level) and (2) implementation techniques (e.g., prompt engineering versus fine-tuning) (Table 5).

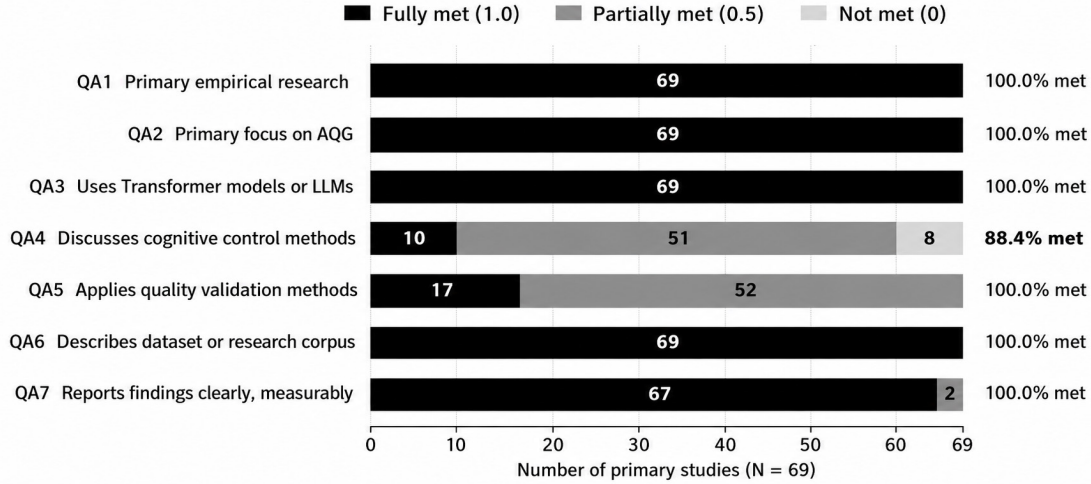


Figure 6. Quality assessment (QA) results for the 69 selected primary studies

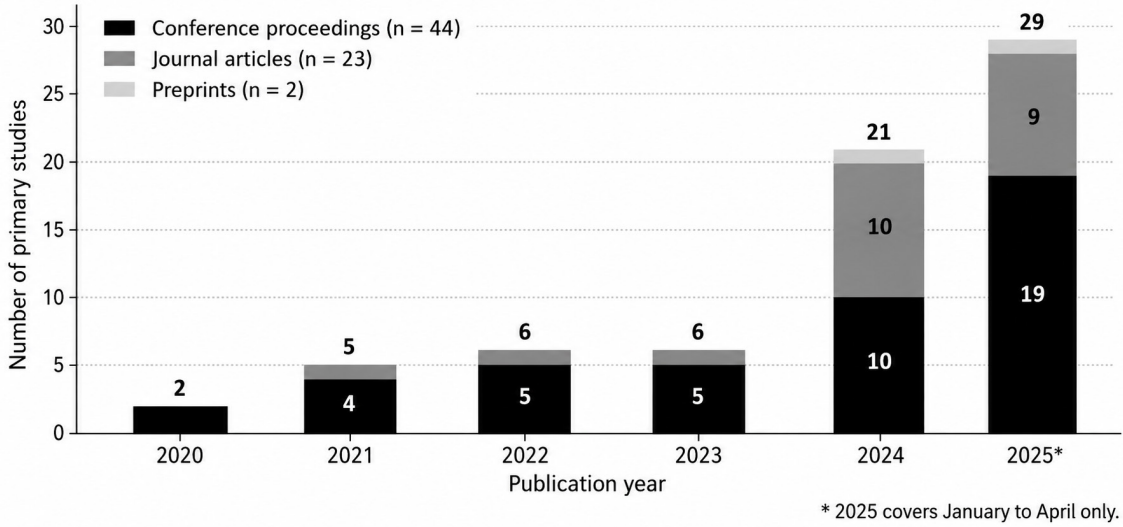


Figure 7. Annual publication counts for the 69 primary studies

The most important finding is the adoption of Bloom’s Taxonomy as the de facto pedagogical framework for cognitive control (Jadhav et al., 2025; Timosh et al., 2025). The majority of studies focusing on control explicitly attempt to guide the model to generate questions at specific cognitive levels (e.g., C1–C6) (Scaria et al., 2024). Conversely, a minority of studies use the simpler difficulty-level framework (e.g., easy, medium, hard) (Shah et al., 2024) or advanced psychometric models such as IRT (Tomikawa & Uto, 2024).

To implement this cognitive control, and Bloom’s Taxonomy in particular, we identified two main streams of techniques, which often correlate with the type of model used. The first stream is prompt engineering for large-scale LLMs (such as the GPT series models). This approach (Lee et al., 2024;

Wu et al., 2023) does not require fine-tuning and relies on instructions within the prompt (e.g., zero-shot or few-shot) (Lee et al., 2024). Studies specifically design complex prompts to direct the LLM to generate Bloom-based questions (L. Wang et al., 2025). The second stream is fine-tuning of encoder-decoder models such as T5 and mT5. This approach involves fine-tuning the model on a dataset explicitly labelled with Bloom levels (Atmajaya et al., 2025; Timosh et al., 2025). This approach is often considered to yield more precise control. Several recent studies have also explored parameter-efficient fine-tuning (PEFT) methods, such as LoRA, to achieve Bloom-based control efficiently (Lamsiyah et al., 2024).

Table 5. Taxonomy of AQG control methods

Implementation method	Cognitive control framework	Implementation description	Common leading models	Example primary studies
1. Prompt engineering (Stream B: flexibility)	Bloom's Taxonomy (dominant); difficulty level (common)	Uses in-context instructions and few-shot examples to guide LLMs in generating questions at specified cognitive levels. Zero-shot and few-shot prompting: direct instructions, with or without a small number of examples, guide the generated output (Lee et al., 2024). Structured prompting: more complex prompts (e.g., with personas) are designed to improve output quality (L. Wang et al., 2025; Wu et al., 2023).	GPT-series models (GPT-3.5, GPT-4), Llama, and chat models	(Lee et al., 2024; L. Wang et al., 2025)
2. Fine-tuning (Stream A: precision)	Bloom's Taxonomy (dominant); Item Response Theory (IRT)	Trains or adjusts model parameters (full fine-tuning or PEFT/LoRA) on an explicitly labelled dataset to achieve high-precision control. Full fine-tuning: encoder-decoder models (T5/mT5) are retrained on Bloom-labelled data (Atmajaya et al., 2025). PEFT (e.g., LoRA): parameters are adjusted efficiently for Bloom-level control (Lamsiyah et al., 2024).	T5, mT5, BART (encoder-decoder)	(Atmajaya et al., 2025; Timosh et al., 2025)
3. Other hybrid methods	Difficulty level (general)	Uses reinforcement learning (RL) or psychometric models such as IRT to calibrate output difficulty (easy, medium, hard) instead of fixed Bloom levels.	Various	(J. Feng & Mostow, 2021; Shah et al., 2024; Tomikawa & Uto, 2024)

Two patterns emerge from this mapping. First, studies that use fine-tuning (T5) tend to build highly specialised systems for specific domains such as programming or cybersecurity (Mulla & Gharpure, 2023; Suhartono et al., 2024). Second, studies that use prompt engineering (GPT) emphasise flexibility and zero-shot performance across domains (Lee et al., 2024; Wu et al., 2023). Choosing between these two streams, precise fine-tuning and flexible prompting, is a central design decision in current automatic question generation research (Atmajaya et al., 2025; Lamsiyah et al., 2024).

RQ1: LANDSCAPE OF AQG RESEARCH FOR LOW-RESOURCE LANGUAGES

The 69 studies are unevenly distributed across languages. Most of them (60 of 69, 87.0%) address English exclusively, leaving 9 studies (13.0%) that address other languages. Within that remainder, Indonesian is the most frequently studied language ($n = 5$, 7.2%) (Atmajaya et al., 2025; Awalurrahman & Budi, 2024; Fuadi & Wibawa, 2022; Suhartono et al., 2024; Vincentio & Suhartono, 2022). The remaining studies address Spanish ($n = 2$, 2.9%), Thai ($n = 1$, 1.4%), and Vietnamese ($n = 1$, 1.4%) (de-Fitero-Dominguez et al., 2024; D. H. Nguyen et al., 2025; Wiwatbutsiri et al., 2022). The results for RQ1, therefore, show a pronounced linguistic imbalance in current AQG research: datasets and models are concentrated on English, while low-resource languages are only marginally represented.

RQ2: CURRENT ADVANCED METHODS FOR COGNITIVE CONTROL

Across the 69 included studies, cognitive control mechanisms fall into two methodological streams, as shown in Figure 8. Stream A (precision), employed in 22 studies (31.9%), uses fine-tuning approaches such as PEFT, commonly applied to encoder–decoder models, to achieve strict control over Bloom’s Taxonomy (Atmajaya et al., 2025; Suhartono et al., 2024). In contrast, Stream B (flexibility), adopted in 37 studies (53.6%), uses prompt engineering to exploit the zero-shot and few-shot capabilities of general-purpose LLMs (Awalurrahman & Budi, 2024; Lee et al., 2024).

A further 5 studies (7.2%) employ hybrid approaches that combine both streams, and 5 studies (7.2%) use alternative mechanisms, such as IRT or complexity-based control. Regarding model architecture, LLM/GPT-based models dominate the field, used in 41 studies (59.4%), while T5 and other encoder–decoder models remain relevant in 20 studies (29.0%); the remaining 8 studies (11.6%) combine both architectures within a single pipeline or use other specialised models, particularly for fine-tuning-based cognitive control in low-resource language settings.

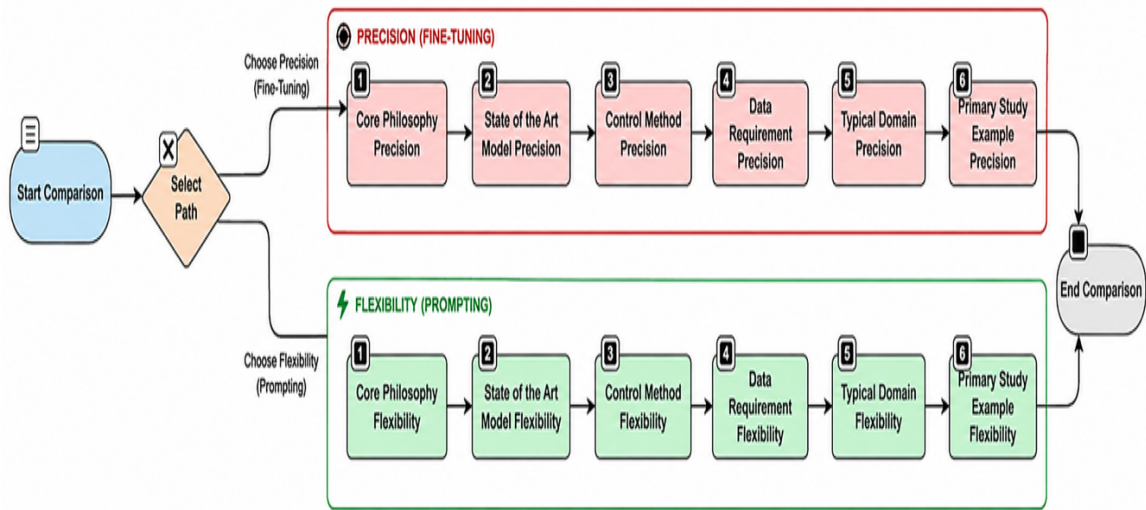


Figure 8. Precision-oriented fine-tuning (Stream A) versus flexibility-oriented prompting (Stream B)

Within Stream B, prompting methods include zero-shot and few-shot prompting, in which the instruction, with or without worked examples, specifies the intended cognitive level (Lee et al., 2024; Wu et al., 2023). Other studies use more elaborate strategies such as persona prompting or chain-of-thought (CoT) prompting (L. Wang et al., 2025). Prompts are also augmented with external sources (e.g., PDF documents) through retrieval-augmented generation (RAG) (Maity et al., 2025b; Mary et al., 2025). Table 6 summarises these prompting types.

Table 6. Taxonomy of prompt engineering techniques for cognitive control (Stream B)

Prompting type (Stream B: flexibility)	Description	Generic prompt template or example	Example primary studies
1. Zero-shot (direct instruction)	Relies on the in-context learning ability of the LLM. The target cognitive level is stated directly in the instruction, without examples.	<i>“Generate a C4-Analysis multiple-choice question based on the following context: [Context ...]”</i>	(Zichao Wang et al., 2022; Wu et al., 2023)
2. Few-shot (with examples)	Supplies one to five worked examples so that the model can infer both the required format and the target cognitive level. Reported as particularly effective for low-resource languages.	<i>“Context: [Context A], Q-C2: [Question A] Context: [Context B], Q-C2: [Question B] Context: [Context C], Q-C2: ...”</i>	(Awalurahman & Budi, 2024; Lee et al., 2024)
3. Persona and CoT (role-playing instruction)	Instructs the model to reason step by step or adopt an expert role to improve output quality and adherence to the target level. Produces longer and more elaborate prompts.	<i>“You are a pedagogy expert specialising in Bloom’s Taxonomy. Your task is to create one ‘C4-Analysis’ question. Context: [Context ...] Question:”</i>	(Shah et al., 2024; L. Wang et al., 2025; Xu et al., 2024)
4. Retrieval-augmented generation (RAG)	Inserts context retrieved from source documents directly into the prompt to limit hallucination and to ground the generated item in the source material.	<i>“Using ONLY the context provided below, generate a C1-Remembering question. Retrieved Context: [Text chunk from PDF ...] Question:”</i>	(Maity et al., 2025b; Mary et al., 2025)

Beyond Bloom’s Taxonomy, emerging control mechanisms offer greater granularity. W. Feng et al. (2025) and L. Yang et al. (2025) introduced complexity-controllable generation using reasoning-augmented sampling and definition constraints. Z. Li et al. (2025) and Z. Yang et al. (2024) addressed topic control scalability. Hybrid approaches further combine RAG with custom knowledge bases (Sayed et al., 2025) and with soft-prompting techniques (Ma et al., 2025). The results for RQ2 show

that cognitive control in AQG is strategically divided between precision-based fine-tuning and flexibility-based prompting, with emerging mechanisms offering increasingly granular control over complexity and topic.

RQ3: QUALITY VALIDATION METHODS FOR PEDAGOGICAL QUALITY

Quality assessment in AQG is reported in all 69 studies, with no study relying solely on automated n-gram metrics. Automated metrics such as ROUGE and BLEU are used to measure grammatical accuracy (Steuer et al., 2021; H.-C. Wang et al., 2024). Newer automated metrics, such as MIRROR (Deroy et al., 2024), are also used to evaluate the outputs of both LLM-based and rule-based systems (Lodovico Molina et al., 2024; Moore et al., 2023).

Regarding human evaluation, full docimological rigour in SME assessment is reported in 17 studies (24.6%) (Van Campenhout et al., 2021; Wróblewska et al., 2025). The remaining 52 studies (75.4%) employ a limited or mixed form of human assessment alongside automated metrics. Manual evaluations typically assess educational reliability, factual accuracy, distractor quality, and docimological validity (Elkins et al., 2023; Grévisse et al., 2024), as well as the alignment of item difficulty with Bloom’s Taxonomy (Scaria et al., 2024). Table 7 summarises the manual evaluation metrics reported across the corpus.

Table 7. Manual AQG evaluation metrics

Metric type	Description	Example primary studies
Baseline criteria		
1. Fluency and grammar	Assesses whether the generated question and answer are written with correct grammar, clarity, and comprehensibility (linguistic quality).	(Contreras-Arguello et al., 2025; Elkins et al., 2023; Mulla & Gharpure, 2023; Steuer et al., 2021; Van Campenhout et al., 2021)
2. Relevance	Assesses whether the generated question and answer are relevant to the provided content/context (topical alignment).	(Contreras-Arguello et al., 2025; Gomazkova et al., 2024; H. A. Nguyen et al., 2022; Saha, 2021; Van Campenhout et al., 2021; Zichao Wang et al., 2022)
Pedagogical criteria		
3. Factual accuracy	Assesses whether the information in the question and the key answer is factually accurate with respect to the source context. Critical in high-stakes domains such as medicine and law.	(Deroy et al., 2025; Elkins et al., 2023; Guo et al., 2025; Kauchak et al., 2024; Maity et al., 2025a)
4. Distractor quality	Assesses whether the distractors in MCQs are (a) relevant to the topic, (b) incorrect but plausible, and (c) not easily eliminated by guessing.	(Atmajaya et al., 2025; Kauchak et al., 2024; Shwe et al., 2024; Winata et al., 2025; Yin & Li, 2025)
5. Docimological and pedagogical quality	A holistic metric assessing whether the item is effective as a testing instrument. It evaluates issues such as ambiguity, fairness, and misleading content.	(Elkins et al., 2023; Grévisse et al., 2024; Mulla & Gharpure, 2023; Saha, 2021; Zichao Wang et al., 2022; Wróblewska et al., 2025)

Metric type	Description	Example primary studies
6. Bloom’s Taxonomy compliance (central metric for this review)	Assesses whether the cognitive level of the generated question matches the level targeted during generation (e.g., C4, analysing).	(Guo et al., 2025; Jadhav et al., 2025; Lamsiyah et al., 2024; Scaria et al., 2024; Suhartono et al., 2024; Timosh et al., 2025)

In addition to standard manual and automated metrics, the literature reports hybrid validation frameworks. These include training classifiers on SME annotations to predict pedagogical quality and taxonomy alignment (Guo et al., 2025; Saha, 2021) and using autonomous LLM agents to simulate SME evaluations (Levine et al., 2025). Furthermore, IRT-based approaches are used to assess latent difficulty and discriminative ability as tunable parameters before deployment (J. Feng & Mostow, 2021; Tomikawa & Uto, 2024). The results for RQ3 show that quality validation in AQG is moving beyond standard automated metrics towards hybrid frameworks and psychometric approaches that aim to measure pedagogical quality and to scale its assessment.

RQ4: MAIN APPLICATION DOMAINS AND DOMINANT LLM MODELS

Analysis of application domains shows that most studies ($n = 58$, 84.1%) address general education, covering K–12, post-secondary, and general instructional settings. Studies devoted to particular domains are far fewer: medical education ($n = 1$, 1.4%), computer science and programming ($n = 1$, 1.4%), and law ($n = 1$, 1.4%) (Deroy et al., 2025; Kauchak et al., 2024; Mulla & Gharpure, 2023). Studies in STEM fields cover mathematics ($n = 2$, 2.9%) (Hwang & Utami, 2024; R. Li et al., 2024) and geographic information systems ($n = 1$, 1.4%) (Zhiyun Wang et al., 2025). The remaining 5 studies (7.2%) did not specify an application domain.

Two types of model architecture predominate. The first comprises smaller models based on the T5 architecture, including the multilingual mT5, which appear in 20 studies (29.0%) (Awalurahman & Budi, 2024; Suhartono et al., 2024). The second comprises large models from the GPT family, which appear in 41 studies (59.4%) (Lee et al., 2024). The remaining 8 studies (11.6%) show both types of architecture in a single pipeline.

Studies published in 2025 account for 29 of the 69 studies (42.0%). In this subset, Asri et al. (2025) describe implementations in the higher education systems of Morocco and Switzerland. Arunnehr et al. (2025) and Snegaa et al. (2025) examine AQG configurations for intelligent assistants, addressing learning analytics and roadmap generation, respectively. Poonkuzhali et al. (2025) and Bernard and Graf (2025) propose integrated educational systems for personalised learning. The results for RQ4 show that although AQG applications are heavily concentrated in general education and rely on large models, recent work indicates a clear shift towards integrating these tools into complete, personalised educational systems.

THE PROPOSED PEDAGOGICAL ALIGNMENT FRAMEWORK

1. Cognitive targeting. The process begins with the curriculum: educators specify the intended learning goal and map it to a level of Bloom’s Taxonomy, for example, C4 (analysing). This specification sets the cognitive target that the control strategy then uses to tune the AQG system towards specific educational outcomes rather than towards factoid questions of little educational value (Jadhav et al., 2025; Scaria et al., 2024; Timosh et al., 2025).

2. Context curation. The quality of the generated items depends directly on the quality of the source material. This stage involves compiling foundational educational materials (textbooks, lecture transcripts) to serve as the knowledge base. Selecting culturally relevant, information-dense texts for fine-

tuning helps to relieve data scarcity and is one of the most important stages for low-resource languages such as Indonesian (Fuadi & Wibawa, 2022; Winata et al., 2025).

3. Controlled generation. With the intended cognitive objectives (step 1) and the selected context (step 2), the system determines which generation mechanism to deploy. For high-precision cases (such as C3, applying), Stream A (fine-tuning with PEFT) is used, drawing on T5 or locally adapted models such as IndoBART (Atmajaya et al., 2025; Lamsiyah et al., 2024; Suhartono et al., 2024). For Stream B cases (more open, exploratory questions), prompt engineering with LLMs is applied (Lee et al., 2024; L. Wang et al., 2025).

4. Dual validation. The framework specifies a two-step validation process designed to relieve the validation bottleneck. The first step is automated screening. An automated assessor, such as a classifier trained on expert annotations, discards candidate items that are irrelevant to the source context or grammatically ill-formed (Mulla & Gharpure, 2023; Saha, 2021). The second step is expert confirmation. One or more SMEs assess the docimological quality of the surviving items, evaluating whether each question is at the intended level of cognitive demand and whether its distractors are plausible (Grévisse et al., 2024; Wróblewska et al., 2025).

5. Integration and delivery. Validated items are not merely stored but integrated into the learning environment, such as a learning management system (LMS) or an intelligent tutoring system (ITS). This stage concerns the delivery format, whether as formative assessments that provide immediate feedback or as CAT items that adjust to learner proficiency (Susanti et al., 2020; Van Campenhout et al., 2021).

6. Quality assurance. Response data from learners is used to flag items that fail to discriminate or exhibit anomalous difficulty patterns, thereby supporting continuous iteration and sustained quality assurance. Tomikawa and Uto (2024) show how this feedback loop turns an AQG system from a static generator into a self-improving one; comparable feedback mechanisms are described by Arunehru et al. (2025). The complete six-stage framework is synthesised in Table 8.

Table 8. Stages of the proposed pedagogical alignment framework

Stage	Actor and method	Goal	Output	Example instruction or action
1. Cognitive targeting	Educators and curriculum designers	Define specific learning goals and map them to cognitive levels (e.g., Bloom’s Taxonomy).	Cognitive target specification (e.g., a C4 analysing objective)	<i>“Define learning objective: Students must analyse the causes of the Java War using C4-Analysis level questions.”</i>
2. Context curation	Subject matter experts (SMEs)	Select authoritative and culturally relevant educational content (textbooks, transcripts) as the knowledge source.	Curated context corpus (high-quality input text)	<i>“Select ‘Chapter 5: Independence Struggle’ from the official Ministry of Education history textbook as the source context.”</i>
3. Controlled generation	Stream A (T5 + PEFT) or Stream B (LLM prompting)	Generate question candidates that strictly adhere to the defined cognitive targets and context.	Candidate question set (raw generated items)	<i>Stream B Prompt: “Act as a pedagogy expert. Generate one multiple-choice question based on the provided text that tests C4-Analysis skills. Ensure distractors are plausible.”</i>

Stage	Actor and method	Goal	Output	Example instruction or action
4. Dual validation	Automated classifier (BERT-based) and human SMEs	Filter grammatical errors computationally and verify docimological quality and fairness pedagogically.	Validated item bank (pedagogically sound questions)	<i>Automated: "Classify question #101: Is the cognitive level predicted as C4?" Human: "Rate the distractor quality on a scale of 1–5. Are there any trick questions?"</i>
5. Integration and delivery	LMS and adaptive systems (CAT)	Deploy validated items into the learning environment for formative or adaptive assessment.	Live assessment activity (student interaction)	<i>"Import validated item bank into Moodle/Canvas LMS for the Mid-Term Adaptive Exam session."</i>
6. Quality assurance	Learning analytics and psychometrics (IRT)	Evaluate item performance based on actual student response data and refine the generation process.	Continuous feedback loop (refined model and parameters)	<i>"Calculate Item Discrimination Index and Difficulty Parameter (b) for Question #101 based on student response data."</i>

DISCUSSION

The synthesis of 69 primary studies shows that AQG has shifted from a predominantly technical natural language processing task towards an educational technology problem in which pedagogical alignment is the decisive criterion. Early work in the corpus focused mainly on producing well-formed questions, whereas the current frontier centres on managing their cognitive demand and verifying their educational quality. This section discusses four themes that follow directly from the four research questions: equity of access (RQ1), the pedagogical consequences of cognitive control (RQ2), the validity of automated assessment (RQ3), and the changing role of the teacher (RQ4).

EQUITY AND INCLUSION: BRIDGING THE LINGUISTIC GAP (RQ1)

The distribution reported for RQ1 has an ethical as well as a technical dimension. Sixty of the 69 studies (87.0%) address English exclusively, leaving 13.0% for all other languages and only 7.2% for Indonesian. English benefits from abundant annotated corpora and mature pretrained models (de-Fitero-Dominguez et al., 2024; Johnson et al., 2025), whereas work on low-resource languages remains sparse and is concentrated in well-resourced domains such as medicine and computer science. Speakers of low-resource languages therefore face a compound disadvantage: neither the training data nor the language models required for reliable generation are available to them, and the pools of subject matter experts needed for pedagogical validation are correspondingly thin.

Closing this gap requires locally constructed corpora and culturally situated systems, as demonstrated by Fuadi and Wibawa (2022), Suhartono et al. (2024), and Atmajaya et al. (2025). The same constraint explains the growing interest in smaller language models for resource-constrained settings (Berger et al., 2024). Encoder–decoder models of the T5 family have proved more tractable for localisation and adaptation than larger proprietary models (Awalurahman & Budi, 2024; Suhartono et al., 2024), and parameter-efficient fine-tuning lowers the computational cost of adapting them further. A parallel line of work extends AQG towards personalised assessment for learners with intellectual disabilities

(Poonkuzhali et al., 2025), broadening the scope of the field from linguistic coverage to equitable access more generally and connecting AQG research to the goal of inclusive quality education.

PEDAGOGICAL IMPLICATIONS OF COGNITIVE CONTROL: PRECISION AND FLEXIBILITY (RQ2)

Aligning generation with Bloom’s Taxonomy reframes the task from one of syntactic production to one of cognitive scaffolding. The two streams identified in RQ2 achieve this alignment by different means, and choosing between them entails a trade-off between precision and flexibility.

In the precision stream (Stream A), the cognitive target is embedded during training: encoder–decoder architectures such as T5 and mT5 are fine-tuned so that the output is constrained to a specified level of the taxonomy, for example, distinguishing C3 (applying) from C4 (analysing). This degree of control matters most in high-stakes assessment, where the cognitive level of an item must be specified unambiguously and remain defensible under scrutiny.

In the flexibility stream (Stream B), prompt engineering applied to general-purpose LLMs such as GPT-4 makes cognitive control available without retraining. Chain-of-thought and persona prompting are used to steer models towards higher-order questions, and the literature shows increasing reliance on such descriptive strategies, including instructions that cast the model as a specialist in cognitive-level assessment (L. Wang et al., 2025). Retrieval-augmented generation is added to limit hallucination and to ground generated items in the source material (Maity et al., 2025b).

The theoretical implication is that adaptive cognitive control is a matter of instructional alignment rather than of implementation detail. Where control mechanisms are weak, an AQG system can produce items that are technically well-formed yet fail to elicit the cognitive process the assessment is intended to measure.

ASSESSMENT VALIDITY: BEYOND AUTOMATED METRICS (RQ3)

The validation methods identified in RQ3 support a clear conclusion: standard natural language processing metrics such as BLEU and ROUGE cannot serve as stand-alone measures of educational quality. These metrics capture surface overlap and grammatical well-formedness, but they do not assess factual accuracy, pedagogical purpose, or adherence to Bloom’s Taxonomy (Elkins et al., 2023; Grévisse, 2024). The literature, therefore, exhibits a validity gap: an item can score highly on n-gram similarity yet have little educational value. Concept mapping research offers one route beyond this limitation: Prasetya et al. (2025) show that cognitive fidelity is reflected in the interrelatedness of knowledge structures, suggesting that AQG systems should generate questions that probe conceptual connections rather than isolated facts.

Because automated metrics are insufficient, evaluation by subject-matter experts remains the reference standard. The finding from RQ3 that 52 of the 69 studies (75.4%) rely on limited or partial human assessment indicates that rigorous expert-led evaluation remains the exception rather than the norm. Manual review is time-consuming, but it is currently the only established means of judging distractor plausibility, cueing, fairness, and docimological validity (Elkins et al., 2023; Grévisse et al., 2024).

Two developments respond to this validation bottleneck. The first is psychometric: frameworks based on Item Response Theory estimate the latent difficulty and discrimination of generated items, moving validation from qualitative judgement to estimable parameters that can be specified before deployment (Tomikawa & Uto, 2024). The second is computational: purpose-built evaluators such as MIRROR, developed explicitly as an alternative to n-gram metrics (Deroy et al., 2024), together with autonomous LLM agents that emulate expert review (Levine et al., 2025), suggest a model-based pre-screening layer that filters generated items before a domain specialist sees them. Neither development removes the expert from the process; both seek to use the expert’s scarce time more efficiently.

TEACHER AGENCY AND CO-DESIGN (RQ4)

The application domains and systemic trends identified in RQ4 point to a redefinition of the teacher's role rather than to the replacement of the teacher. Within co-design workflows, teachers take on the role of assessment designer and prompt author. This reading is supported by the RQ4 finding that recent studies increasingly embed AQG within complete educational systems such as learning management systems, rather than treating generation as a stand-alone service. For institutional decision-making and responsible classroom deployment, the implication is concrete: teachers require targeted preparation in prompt design and cognitive alignment if they are to evaluate model output critically.

The concentration of research in general education (84.1%) at the expense of specialised domains also reveals a clear gap. Systems such as GISedu-GPT and domain-specific intelligent learning assistants indicate that AQG performs best when supplied with domain knowledge and a curated knowledge base, and that generic foundation models underperform domain-grounded models on tasks requiring complex reasoning (Zhiyun Wang et al., 2025). Consistent with this, L. Wang et al. (2025) direct attention to teachers' collective pedagogical expertise as a resource for prompt design. Taken together, these findings argue for a human-in-the-loop configuration in which the model drafts and the teacher selects, adapts, and validates. Realising that the configuration will require progress on current limitations in reasoning over non-linear narratives and implicit meaning (Winata et al., 2025), and may benefit from extending AQG to multimodal sources such as video and audio (Gala et al., 2021).

LIMITATIONS

This systematic literature review provides a detailed synthesis of existing control and validation techniques, but several limitations should be noted. First, the corpus is restricted to studies that address cognitive control or quality validation, so work concerned solely with generation architecture falls outside its scope. Second, the review window (2020–2025) covers the post-Transformer period and therefore excludes earlier rule-based and statistical work. Third, although we highlight the gap in Indonesian-language publications, the search strategy relied on English keywords and may therefore have overlooked publications in local languages that are not indexed in global repositories. Fourth, screening was performed by a single reviewer; although every decision was guided by predefined PRISMA criteria and a structured QA rubric, the absence of a second independent reviewer meant that inter-rater agreement statistics could not be calculated, which may introduce a degree of selection bias.

CONCLUSION

This systematic literature review reframes AQG not as a subtask of natural language generation (NLG) but as a distinct field of enquiry, which may be termed pedagogical alignment engineering. Specifically, by synthesising 69 studies published between 2020 and 2025, this review produces four main contributions: (1) empirical evidence of a low-resource language gap, (2) a taxonomy of cognitive control methods, (3) a synthesis of validation practices, and (4) a proposed Pedagogical Alignment Framework. Whereas early work approached the problem as one of linguistic fluency, more recent studies reframe it in terms of Bloom's Taxonomy, which has become the default framework for constraining generated text to a specified cognitive level. This shift is clearly reflected in our data, with 88.4% of the reviewed studies explicitly discussing cognitive control mechanisms and all 69 studies applying quality validation beyond n-gram metrics. However, only 24.6% did so with full docimological rigour.

Summary of key findings. Across the four research questions, the current landscape of AQG reveals significant trends and disparities. First, the field is heavily dominated by English-centric research (87.0%), leaving a substantial gap for low-resource languages such as Indonesian (7.2%). To bridge this gap, smaller fine-tuned models (such as T5 architectures) remain crucial for localisation, even as the field explores the flexibility of general-purpose large language models. Second, application domains are disproportionately concentrated in general education (84.1%), indicating a clear need for domain-specific AQG systems. Finally, although automated NLP metrics remain widely used, they often fail to capture educational validity. Consequently, expert validation remains the indispensable reference standard, yet it is applied with full docimological rigour in only 24.6% of studies, while the remaining 75.4% relying on limited or partial human review. Future frameworks will need to support and scale expert evaluation rather than replace it.

Theoretical contributions. The review identifies a bifurcation between fine-tuned encoder–decoder models, such as T5 and mT5 (Stream A, precision), and prompt engineering for large language models, such as GPT (Stream B, flexibility). This classification links control strategy to architectural choice and makes the trade-off between pedagogical precision and generative flexibility explicit. We further argue that conventional automatic evaluation (e.g., BLEU and ROUGE) is inadequate for educational applications, and we propose expert-assessed docimological quality as the appropriate benchmark.

Methodological contributions. In response to the scarcity of resources for low-resource languages such as Indonesian, this review proposes the Pedagogical Alignment Framework. This methodological contribution proposes a closed-loop approach that combines the zero-shot capabilities of LLMs with parameter-efficient fine-tuning and automated assessors to relieve the validation bottleneck. Further research should pursue this integrated approach rather than incremental technical optimisation, so that AQG systems are not only linguistically adequate but also pedagogically appropriate, psychometrically valid, and accessible.

ACKNOWLEDGEMENTS

This research was funded by the Indonesian Endowment Fund for Education (LPDP) on behalf of the Indonesian Ministry of Higher Education, Science and Technology, and managed under the EQUITIVITY Program (Contract No. 4309/B3/DT.03.08/2025).

REFERENCES

- Agarwal, R., Negi, V., Kalra, A., & Mittal, A. (2022). Deep learning and linguistic feature based automatic multiple choice question generation from text. In R. Bapi, S. Kulkarni, S. Mohalik, & S. Peri (Eds.), *Distributed computing and intelligent technology* (pp. 260–264). Springer. https://doi.org/10.1007/978-3-030-94876-4_18
- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Arunnethru, J., Vishnu, S., Ganapathyappan, K., & Kumar, D. (2025). AI-enhanced learning assistant platform: An advanced system for Q&A generation from provided content, answer evaluation and roadmap generation. In P. D. Sivakumar, R. Ramachandran, C. Pasupathi, & P. Balakrishnan (Eds.), *Computing technologies for sustainable development* (pp. 332–346). Springer. https://doi.org/10.1007/978-3-031-82386-2_26
- Asri, H., Carbonel, H., Julien, J. M., & Rochdi, A. (2025). Automated question generation with AI: Enhancing educational systems using LLMs at Cadi Ayyad University and UniDistance Switzerland. In R. González Vallejo, G. Moukhliiss, E. Schaeffer, & V. Paliktzoglou (Eds.), *The Second International Symposium on Generative AI and Education* (pp. 421–432). Springer. https://doi.org/10.1007/978-3-031-98476-1_33
- Atmajaya, P. A., Romadhony, A., & Ihsan, A. F. (2025). Multiple-choice question generation in low resource Indonesian language with T5. *International Journal of Intelligent Engineering and Systems*, 18(5), 363–381. <https://doi.org/10.22266/ijies2025.0630.26>

- Awalurahman, H. W., & Budi, I. (2024). Paragraph vs sentence in automatic question generation fine-tuning using text-to-text transfer transformer for Bahasa Indonesia. In *2024 10th International Conference on Education and Technology (ICET)* (pp. 155–161). IEEE. <https://doi.org/10.1109/ICET64717.2024.10778465>
- Bahri, A., & Oueslati, W. (2025). Towards a RAG-based framework for accounting practices: Context-aware question generation. In A. Mirzazadeh, Z. Molamohamadi, B. Erdebili, E. Babaee Tirkolaei, & G.-W. Weber (Eds.), *Communications in computer and information science: Vol. 2651. Science, engineering management and information technology* (pp. 42–53). Springer. https://doi.org/10.1007/978-3-032-04225-5_3
- Berger, J., Koß, J., Stamatakis, M., Hoppe, A., Ewerth, R., & Wartena, C. (2024). Question generation capabilities of “small” large language models. In A. Rapp, L. Di Caro, F. Mezziane, & V. Sugumaran (Eds.), *Lecture notes in computer science: Vol. 14763. Natural language processing and information systems* (pp. 183–194). Springer. https://doi.org/10.1007/978-3-031-70242-6_18
- Bernard, J., & Graf, S. (2025). Language models for educational question generation: Practical challenges, personalization opportunities, and parameter optimization. In S. Graf & A. Markos (Eds.), *Lecture notes in computer science: Vol. 15723. Generative systems and intelligent tutoring systems* (pp. 144–158). Springer. https://doi.org/10.1007/978-3-031-98281-1_11
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay Company.
- Bulathwela, S., Muse, H., & Yilmaz, E. (2023). Scalable educational question generation with pre-trained language models. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Lecture notes in computer science: Vol. 13916. Artificial intelligence in education* (pp. 223–228). Springer. https://doi.org/10.1007/978-3-031-36272-9_27
- Contreras-Arguello, M. L., Paredes-Valverde, M. A., Vásquez, A. M. T., & Salas-Zárate, M. P. (2025). Automatic generation of summaries and questions to support the reading comprehension process. In R. Valencia-García, T. Borodulina, J. Del Cioppo-Morstadt, C. E. Moran-Castro, & N. Vera-Lucio (Eds.), *Communications in computer and information science: Vol. 2276. Technologies and innovation* (pp. 81–92). Springer. https://doi.org/10.1007/978-3-031-75702-0_7
- de-Fitero-Domínguez, D., García-Cabot, A., & García-López, E. (2024). Automated multiple-choice question generation in Spanish using neural language models. *Neural Computing and Applications*, 36(29), 18223–18235. <https://doi.org/10.1007/s00521-024-10076-7>
- Deroy, A., Ghosh, K., & Ghosh, S. (2025). Investigating legal question generation using large language models. *Artificial Intelligence and Law*. Advance online publication. <https://doi.org/10.1007/s10506-025-09452-y>
- Deroy, A., Maity, S., & Sarkar, S. (2024). MIRROR: A novel approach for the automated evaluation of open-ended question generation. arXiv. <https://doi.org/10.48550/arXiv.2410.12893>
- Elkins, S., Kochmar, E., Serban, I., & Cheung, J. C. K. (2023). How useful are educational questions generated by large language models? In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 536–547). Springer. https://doi.org/10.1007/978-3-031-36336-8_83
- Feng, J., & Mostow, J. (2021). Towards difficulty controllable selection of next-sentence prediction questions. *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 262–267). International Educational Data Mining Society. https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_262.pdf
- Feng, W., Tran, P., Sireci, S., & Lan, A. S. (2025). Reasoning and sampling-augmented MCQ difficulty prediction via LLMs. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education* (pp. 31–45). Springer. https://doi.org/10.1007/978-3-031-98459-4_3
- Fuadi, M., & Wibawa, A. D. (2022). Automatic question generation from Indonesian texts using text-to-text transformers. *Proceedings of the International Conference on Electrical and Information Technology, Malang, Indonesia*, 84–89. <https://doi.org/10.1109/IEIT56384.2022.9967858>

- Gala, R., Vijayaraghavan, R., Kiwelekar, A., & Nikam, V. (2021). Real-time cognitive evaluation of online learners through automatically generated questions. *Proceedings of the IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing, Banff, Alberta, Canada*, 53–58. <https://doi.org/10.1109/IC-CICC53683.2021.9811326>
- Gomazkova, Y., Sychev, O., & Gumerov, M. (2024). Improving efficiency of natural-language text generation for automatic pedagogical questions. In O. Gervasi, B. Murgante, C. Garau, D. Taniar, A. M. A. C. Rocha, & M. N. Faginas Lago (Eds.), *Computational science and its applications – ICCSA 2024 workshops* (pp. 37–50). Springer. https://doi.org/10.1007/978-3-031-65223-3_3
- Grévisse, C. (2024). Comparative quality analysis of GPT-based multiple choice question generation. In H. Florez & M. Leon (Eds.), *Applied informatics* (pp. 435–447). Springer. https://doi.org/10.1007/978-3-031-46813-1_29
- Grévisse, C., Pavlou, M. A. S., & Schneider, J. G. (2024). Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. *SN Computer Science*, 5(5), Article 636. <https://doi.org/10.1007/s42979-024-02963-6>
- Guo, M., Sun, B., He, J., & Zhang, F. (2025). Research on automatic question generation methods for niche subjects based on large language models. *Proceedings of the 14th International Conference on Educational and Information Technology, Guangzhou, China*, 106–110. <https://doi.org/10.1109/ICEIT64364.2025.10976111>
- Hwang, W.-Y., & Utami, I. Q. (2024). Using GPT and authentic contextual recognition to generate math word problems with difficulty levels. *Education and Information Technologies*, 29(13), 17635–17663. <https://doi.org/10.1007/s10639-024-12537-x>
- Jadhav, P., Harnekar, M., Jalgaonkar, A., More, S., Khandekar, S., & Bhingarde, K. (2025). Integrated framework for Bloom’s based question and short answer generation using transformer. *Proceedings of the International Conference on Soft Computing for Security Applications, Salem, India*, 617–622. <https://doi.org/10.1109/ICSCSA66339.2025.11170844>
- Johnson, B. G., Van Campenhout, R., Jerome, B., Castro, M. F., Bistolfi, R., & Dittel, J. S. (2025). Automatic question generation for Spanish textbooks: Evaluating Spanish questions generated with the parallel construction method. *International Journal of Artificial Intelligence in Education*, 35(3), 1044–1063. <https://doi.org/10.1007/s40593-024-00394-1>
- Kauchak, D., Song, V., Mishra, P., Leroy, G., Harber, P., Rains, S., Hamre, J., & Morgenstein, N. (2024). Automatic generation of a large multiple-choice question-answer corpus. In K. Arai (Ed.), *Intelligent systems and applications* (pp. 55–72). Springer. https://doi.org/10.1007/978-3-031-66428-1_4
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1), 121–204. <https://doi.org/10.1007/s40593-019-00186-y>
- Lamsiyah, S., El Mahdaouy, A., Nourbakhsh, A., & Schommer, C. (2024). Fine-tuning a large language model with reinforcement learning for educational question generation. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (pp. 424–438). Springer. https://doi.org/10.1007/978-3-031-64302-6_30
- Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, 29(9), 11483–11515. <https://doi.org/10.1007/s10639-023-12249-8>
- Levine, J., West, M., & Silva, M. (2025). LLM agents for verifiable question generation and grading. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium, Blue Sky, and WideAIED* (pp. 177–184). Springer. https://doi.org/10.1007/978-3-031-99264-3_22
- Li, R., Wang, Y., Zheng, C., Jiang, Y.-H., & Jiang, B. (2024). Generating contextualized mathematics multiple-choice questions utilizing large language models. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 548–559). Springer. https://doi.org/10.1007/978-3-031-64315-6_48

- Li, Z., Cukurova, M., & Bulathwela, S. (2025). A novel approach to scalable and automatic topic-controlled question generation in education. *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 148–158). Association for Computing Machinery. <https://doi.org/10.1145/3706468.3706487>
- Liao, Y.-H., & Koh, J.-L. (2020). Question generation through transfer learning. In H. Fujita, P. Fournier-Viger, M. Ali, & J. Sasaki (Eds.), *Trends in artificial intelligence theory and applications: Artificial intelligence practices* (pp. 3–17). Springer. https://doi.org/10.1007/978-3-030-55789-8_1
- Lodovico Molina, I., Švábenský, V., Minematsu, T., Chen, L., Okubo, F., & Shimada, A. (2024). Comparison of large language models for generating contextually relevant questions. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruipérez Valiente (Eds.), *Technology enhanced learning for inclusive and equitable quality education* (pp. 137–143). Springer. https://doi.org/10.1007/978-3-031-72312-4_18
- Ma, J., Zhan, R., Wong, D. F., & Chao, L. S. (2025). Activate integrated controllable generation with soft prompt. In D. F. Wong, Z. Wei, & M. Yang (Eds.), *Natural language processing and Chinese computing* (pp. 239–251). Springer. https://doi.org/10.1007/978-981-97-9440-9_19
- Maity, S., Deroy, A., & Sarkar, S. (2024). Investigating large language models for prompt-based open-ended question generation in the technical domain. *SN Computer Science*, 5, Article 1128. <https://doi.org/10.1007/s42979-024-03464-2>
- Maity, S., Deroy, A., & Sarkar, S. (2025a). Can large language models meet the challenge of generating school-level questions? *Computers and Education: Artificial Intelligence*, 8, Article 100370. <https://doi.org/10.1016/j.caeai.2025.100370>
- Maity, S., Deroy, A., & Sarkar, S. (2025b). *Leveraging in-context learning and retrieval-augmented generation for automatic question generation in educational domains*. arXiv. <https://doi.org/10.48550/arXiv.2501.17397>
- Mary, A. C. A., Acharya, P., Rakshinee, R., Jeyaseelan, S., Nandakumar, T., Sakthivel, V., & Prakash, P. (2025). Enhancing question answer generation from PDFs: A fusion of BERT, RAKE, T5 and DistilBERT with RQUGE evaluation. In V. K. Gunjan & J. M. Zurada (Eds.), *Proceedings of the 5th International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications* (pp. 249–260). Springer. https://doi.org/10.1007/978-981-97-8861-3_22
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Moore, S., Nguyen, H. A., Chen, T., & Stamper, J. (2023). Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods. In O. Viberg, I. Jivet, P. Muñoz-Merino, M. Perifanou, & T. Papathoma (Eds.), *Responsive and sustainable educational futures* (pp. 234–248). Springer. https://doi.org/10.1007/978-3-031-42682-7_16
- Mulla, N., & Gharpure, P. (2023). Leveraging well-formedness and cognitive level classifiers for automatic question generation on Java technical passages using T5 transformer. *International Journal of Information Technology*, 15(4), 1961–1973. <https://doi.org/10.1007/s41870-023-01262-2>
- Navas, G., Proaño-Orellana, J., Orizondo, R., & Terreros, A. (2025). Exams as a service: Synergies between ChatGPT, and cloud computing for education. In F. R. Narváez, M. N. Villa, & G. M. Díaz (Eds.), *Smart technologies, systems and applications* (pp. 178–192). Springer. https://doi.org/10.1007/978-3-031-98287-3_13
- Nguyen, D. H., Tran, C., & Tran, H. T. (2025). Enhancing Vietnamese automatic question generation with synthetic data and preference optimization. *Proceedings of the 7th International Conference on Natural Language Processing, Guangzhou, China*, 183–188 <https://doi.org/10.1109/ICNLP65360.2025.11108542>
- Nguyen, H. A., Bhat, S., Moore, S., Bier, N., & Stamper, J. (2022). Towards generalized methods for automatic question generation in educational domains. In I. Hilliger, P. J. Muñoz-Merino, T. De Laet, A. Ortega-Arnanz, & T. Farrell (Eds.), *Educating for a new future: Making sense of technology-enhanced learning adoption* (pp. 272–284). Springer. https://doi.org/10.1007/978-3-031-16290-9_20
- Pan, L., Lei, W., Chua, T.-S., & Kan, M.-Y. (2019). *Recent advances in neural question generation*. arXiv. <https://doi.org/10.48550/arXiv.1905.08949>

- Poonkuzhali, S., Archana, J. S., Naresh, A., & Raj, R. N. (2025). GenAI-powered adaptive assessments for livelihood skills in individuals with intellectual disabilities. In A. Joshi, R. Ragel, M. Mahmud, & S. Kartik (Eds.), *ICT: Applications and social interfaces* (pp. 127–136). Springer. https://doi.org/10.1007/978-981-96-5754-4_12
- Prasetya, D. D., Widiyaningtyas, T., & Hirashima, T. (2025). Interrelatedness patterns of knowledge representation in extension concept mapping. *Research and Practice in Technology Enhanced Learning*, 20, Article 9. <https://doi.org/10.58459/rptel.2025.20009>
- Saha, S. K. (2021). Towards development of a system for automatic assessment of the quality of a question paper. *Smart Learning Environments*, 8, Article 4. <https://doi.org/10.1186/s40561-021-00148-9>
- Sayed, A., Kanojia, M., & Nabajja, S. (2025). Subjective question bank generation using large language models with custom knowledge base. In V. Sakalauskas, A. Bajaj, A. Abraham, K. R. Madhavi, & P. Manghirmalani Mishra (Eds.), *Bio-inspired computing* (pp. 250–259). Springer. https://doi.org/10.1007/978-3-031-78946-5_24
- Scaria, N., Chenna, S. D., & Subramani, D. (2024). Automated educational question generation at different Bloom’s skill levels using large language models: Strategies and evaluation. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (pp. 158–172). Springer. https://doi.org/10.1007/978-3-031-64299-9_12
- Shah, A., Devmane, A., Ranka, M., & Churi, P. (2024). Improvised progressive model based on automatic calibration of difficulty level: A practical solution of competitive-based examination. *Education and Information Technologies*, 29(6), 6909–6946. <https://doi.org/10.1007/s10639-023-12045-4>
- Shwe, L. L., Matayong, S., & Witosurapot, S. (2024). The unified difficulty ranking mechanism for automatic multiple choice question generation in digital storytelling domain. *Education and Information Technologies*, 29, 20317–20350. <https://doi.org/10.1007/s10639-024-12666-3>
- Snegaa, A., Sabiyath Fatima, N., Karthiga, I., Amsavalli, S., Nazreen, A., & Jhasim Hassan, J. (2025). Advanced learning system — Automatic MCQ generator with answer evaluation and performance analysis using LLM. In A. E. Hassanien, S. Anand, A. Jaiswal, & P. Kumar (Eds.), *Innovative computing and communications* (pp. 607–622). Springer. https://doi.org/10.1007/978-981-96-6883-0_42
- Steuer, T., Bongard, L., Uhlig, J., & Zimmer, G. (2021). On the linguistic and pedagogical quality of automatic question generation via neural machine translation. In T. De Laet, R. Klemke, C. Alario-Hoyos, I. Hilliger, & A. Ortega-Arranz (Eds.), *Technology-enhanced learning for a free, safe, and sustainable world* (pp. 289–294). Springer. https://doi.org/10.1007/978-3-030-86436-1_22
- Suhartono, D., Majiid, M. R. N., & Fredyan, R. (2024). Towards automatic question generation using pre-trained model in academic field for Bahasa Indonesia. *Education and Information Technologies*, 29(16), 21295–21330. <https://doi.org/10.1007/s10639-024-12717-9>
- Susanti, Y., Tokunaga, T., & Nishikawa, H. (2020). Integrating automatic question generation with computerised adaptive test. *Research and Practice in Technology Enhanced Learning*, 15, Article 9. <https://doi.org/10.1186/s41039-020-00132-w>
- Timosh, R. R., Joy, E., Evangeline, S., & Aroma, R. J. (2025). Automated question generation with fine-tuned NLP model based on revised Bloom’s taxonomy. *SN Computer Science*, 6, Article 919. <https://doi.org/10.1007/s42979-025-04453-9>
- Tomikawa, Y., & Uto, M. (2024). Difficulty-controllable multiple-choice question generation for reading comprehension using item response theory. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 312–326). Springer. https://doi.org/10.1007/978-3-031-64315-6_27
- Van Campenhout, R., Dittel, J. S., Jerome, B., & Johnson, B. G. (2021). Transforming textbooks into learning by doing environments: An evaluation of textbook-based automatic question generation. *Proceedings of the Third Workshop on Intelligent Textbooks* (CEUR Workshop Proceedings, Vol. 2895, pp. 60–73). CEUR-WS. <https://ceur-ws.org/Vol-2895/paper06.pdf>

- Vincentio, K., & Suhartono, D. (2022). Automatic question generation using RNN-based and pre-trained transformer-based models in low resource Indonesian language. *Informatica*, 46(7), 103–118. <https://doi.org/10.31449/inf.v46i7.4236>
- Wang, H.-C., Chiang, Y.-H., & Chen, I.-F. (2024). A method for generating course test questions based on natural language processing and deep learning. *Education and Information Technologies*, 29(7), 8843–8865. <https://doi.org/10.1007/s10639-023-12159-9>
- Wang, L., Song, R., Guo, W., & Yang, H. (2025). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 33(3), 2559–2584. <https://doi.org/10.1080/10494820.2024.2412082>
- Wang, Z. [Zhiyun], Zhang, Y., Min, W., Guan, Q., & Yu, W. (2025). GISedu-GPT: A large language model framework with prior knowledge for GIS education question bank generation. *Journal of Geography in Higher Education*, 50(1), 72–99. <https://doi.org/10.1080/03098265.2025.2549784>
- Wang, Z. [Zichao], Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 153–166). Springer. https://doi.org/10.1007/978-3-031-11644-5_13
- Winata, A. J., Surjawati, D. J., & Mawardi, V. C. (2025). Utilizing large language models for developing automatic question generation in education. *Proceedings of the International Conference on Advancement in Data Science, E-learning and Information System, Bandung, Indonesia*, 1–6. <https://doi.org/10.1109/ICADEIS65852.2025.10933227>
- Wiwatbutsiri, N., Suchato, A., Punyabukkana, P., & Tuaycharoen, N. (2022). Question generation in the Thai language using MT5. *Proceedings of the 19th International Joint Conference on Computer Science and Software Engineering, Bangkok, Thailand*, 1–6. <https://doi.org/10.1109/IJCSSE54890.2022.9836271>
- Wróblewska, A., Grabek, B., Świstak, J., & Dan, D. (2025). Evaluating LLM-generated Q&A test: A student-centered study. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education* (pp. 277–289). Springer. https://doi.org/10.1007/978-3-031-98417-4_20
- Wu, Y., Nouri, J., Megyesi, B., Henriksson, A., Duneld, M., & Li, X. (2023). Towards data-effective educational question generation with prompt-based learning. In K. Arai (Ed.), *Intelligent computing* (pp. 161–174). Springer. https://doi.org/10.1007/978-3-031-37717-4_11
- Xu, J., Zhang, L., Sun, Y., Wang, J., & Duan, S. (2024). Personalized questioning teaching mode for English reading in junior high school based on automatic question generation. In J. Gan, Y. Pan, J. Zhou, D. Liu, X. Song, & Z. Lu (Eds.), *Computer science and educational informatization* (pp. 90–102). Springer. https://doi.org/10.1007/978-981-99-9492-2_9
- Yang, L., Yuan, J., Kong, C., Yu, J., Chong, R., Liu, Z., & Yang, E. (2025). Tailored definitions with easy reach: Complexity-controllable definition generation. *IEEE Transactions on Big Data*, 11(4), 2061–2071. <https://doi.org/10.1109/TBDDATA.2024.3522805>
- Yang, Z., Jiang, H., Deng, A., & Li, Y. (2024). Topic-oriented controlled text generation for social networks. *Journal of Signal Processing Systems*, 96(2), 131–151. <https://doi.org/10.1007/s11265-023-01907-2>
- Yin, Z., & Li, L. (2025). KPQDG: A key sentence prompt-tuning for automatic question-distractor pairs generation. *Neural Computing and Applications*, 37(18), 11843–11861. <https://doi.org/10.1007/s00521-025-11149-x>

APPENDIX. QUALITY ASSESSMENT PROTOCOL

A.1 Full QA Rubric with Scoring Criteria

This appendix presents the complete quality assessment (QA) rubric applied to evaluate all candidate studies during the systematic review process. The rubric consists of seven criteria divided into two tiers: three mandatory criteria (QA1–QA3) and four quality criteria (QA4–QA7). A study is included in the final corpus ($n = 69$) if and only if it achieves a perfect score on all mandatory criteria and a cumulative score greater than 2.0 out of 4.0 on the quality criteria.

A.2 Scoring Legend

Table A1 presents the scoring scale used across all QA criteria.

Table A1. Scoring legend

Score	Label	Interpretation
1.0	Yes — meets fully	The study fully meets the criterion as described in the rubric.
0.5	Partially — meets partly	The study partially meets the criterion — relevant content is present but incomplete, implicit, or lacks methodological rigour.
0	No — does not meet	The study does not meet the criterion. The relevant dimension is absent or contradicted.

A.3 Quality Assessment Rubric (QA1–QA7)

Table A2 presents the full rubric with operational definitions for each scoring level (Yes = 1, Partially = 0.5, No = 0).

Table A2. Full quality assessment rubric with operational scoring definitions

QA criteria	Category	Question/description	Yes (1.0) meets fully	Partially (0.5) meets partly	No (0) does not meet	Weight
Mandatory Criteria (QA1–QA3): A study must score 1 on all three to be considered for further evaluation						
QA1	Study type	Is the study a primary empirical research paper (journal article, conference paper, or peer-reviewed proceedings)?	The paper reports original experiments, evaluations, or implementations with explicit methodology, corpus, and measurable results.	— (No partial score; a study either is or is not primary research.)	Secondary study (review, survey, opinion, or position paper) without experimental data.	Mandatory
QA2	Focus on AQG	Is the primary focus of the study on automatic question generation (AQG) — specifically the generation of questions from text for educational or assessment purposes?	AQG is the main contribution, evaluation target, or research problem. The paper’s core system generates questions from text.	— (No partial score; AQG must be the primary, not incidental, focus.)	AQG is incidental or secondary. The paper focuses on QA, text summarisation, NLG in general, or other NLP tasks where question generation is not the primary contribution.	Mandatory
QA3	Model architecture	Does the study employ a Transformer-based pre-trained	Uses Transformer-based PLM (e.g., T5, BERT, mT5,	— (No partial score; the Trans-	Uses only rule-based systems, template-based approaches,	Mandatory

QA criteria	Category	Question/description	Yes (1.0) meets fully	Partially (0.5) meets partly	No (0) does not meet	Weight
		language model (PLM) or a large language model (LLM) as its core generation architecture?	BART) or an LLM (e.g., GPT-2/3/4, LLaMA, ChatGPT, PaLM, Gemini) explicitly named and applied to AQQ.	former/LLM condition must be clearly met.)	pre-Transformer neural models (e.g., seq2seq LSTM), or keyword extraction without any Transformer component.	
Quality Criteria (QA4–QA7): A study must achieve a cumulative score >2.0 out of 4.0 on these criteria. Each criterion uses a 3-point scale: Yes (1), Partially (0.5), No (0).						
QA4	Cognitive control	Does the study explicitly discuss or implement methods for controlling the cognitive demand of generated questions — using a formal framework such as Bloom’s Taxonomy, Item Response Theory (IRT), or explicit difficulty levels?	The paper explicitly names Bloom’s Taxonomy (C1–C6), IRT, or a comparable cognitive framework and demonstrates its application in generating or evaluating questions at specific cognitive levels. Includes both prompt-based control and fine-tuning-based control.	The paper mentions difficulty levels (e.g., easy/medium/hard) or discusses the need for cognitive control without applying a named formal framework. Includes papers that use implicit difficulty proxies without mapping to a formal taxonomy.	The paper does not address cognitive demand, difficulty, or pedagogical level in any systematic way. Questions are generated without any mechanism for controlling their cognitive complexity.	Quality (max 4.0)
QA5	Quality validation	Does the study apply quality validation methods that go beyond standard automated n-gram matching metrics (BLEU, ROUGE, METEOR) to assess the pedagogical or docimological quality of generated questions?	Uses at least one of: (a) human/SME evaluation; (b) assessment validity criteria; (c) docimological quality rubrics; (d) psychometric models (IRT difficulty, discrimination); (e) automated quality classifiers trained on pedagogical annotations (e.g., MIRROR, Bloom classifiers).	Conducts a limited human evaluation without a formal rubric or inter-rater reliability check, or uses automated metrics alongside a brief qualitative discussion — but does not employ a structured pedagogical evaluation framework.	Evaluates generated questions exclusively using automated n-gram metrics (BLEU, ROUGE, METEOR, NLTK) without any human evaluation, expert review, or pedagogically-grounded quality assessment.	Quality (max 4.0)
QA6	Corpus / Dataset	Does the study clearly describe the dataset or corpus	The paper explicitly names the dataset, specifies its	The paper mentions the dataset	The paper does not clearly describe	Quality (max 4.0)

QA criteria	Category	Question/description	Yes (1.0) meets fully	Partially (0.5) meets partly	No (0) does not meet	Weight
		used for training, fine-tuning, and/or evaluation, including its source, size, and domain?	domain and size (number of passages, question-answer pairs), and describes how it was obtained or constructed.	but provides incomplete information — e.g., names the source but does not specify its size, or describes the domain without detailing how the corpus was constructed.	the corpus or dataset. The source of training or evaluation data is absent, ambiguous, or only vaguely referenced.	
QA7	Experimental findings	Does the study present its experimental results clearly and measurably, with sufficient detail to understand the performance of the proposed AQG system?	Results are presented with quantitative metrics. Findings are clearly linked to the research questions. Tables or figures present data in a structured, reproducible manner.	Findings are presented but with incomplete quantitative support — e.g., some metrics are reported but without baselines or comparisons, or qualitative descriptions replace quantitative measurements. Preprint papers may fall here if data is preliminary.	Results are absent, overly vague, or consist solely of descriptive text without any quantitative evaluation.	Quality (max 4.0)
<i>Inclusion threshold:</i> A study is included in the final corpus ($n = 69$) if and only if it: (1) scores 1 on all three mandatory criteria (QA1–QA3), and (2) achieves a cumulative score of >2.0 out of 4.0 on the quality criteria (QA4–QA7). The threshold of 2.0 was set to ensure that each included study demonstrates at least a substantive partial engagement with both cognitive control and pedagogical validation — the two central pillars of this review. Studies that scored ≤ 2.0 on QA4–QA7 ($n = 10$) were excluded because, while they focused on AQG with Transformer/LLM architectures, they lacked sufficient engagement with pedagogical quality beyond basic automated metrics.						

A.4 Supplementary Table: QA Scores for All 69 Included Studies

Table A3 presents the quality assessment (QA) scores for each of the 69 studies included in this systematic review. Studies are listed by their corpus identifier (Nos. 1–69). Indonesian-language studies are highlighted in bold. Score legend: 1 = Yes; 0.5 = Partially; 0 = No.

Table A3. QA scores for all 69 included studies — supplementary material

No	Citation	Year	Type	Title (abbreviated)	QA1	QA2	QA3	QA4	QA5	QA6	QA7	Total
Mandatory Criteria (all must = 1 for inclusion): QA1–QA3 scored 0 or 1. Quality Criteria: QA4–QA7 scored 0 / 0.5 / 1 — must total >2.0. Max = 7.0												
1	Agarwal et al. (2022)	2022	Conf. Proc.	Deep learning and linguistic feature based ...	1 ✓	1 ✓	1 ✓	0 ✗	0.5 ~	1 ✓	1 ✓	5.5
2	Arunnehr et al. (2025)	2025	Conf. Proc.	AI-enhanced learning assistant platform: An ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
3	Asri et al. (2025)	2025	Conf. Proc.	Automated question generation with AI ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
4	Atmajaya et al. (2025)	2025	Journal	Multiple-choice question generation in low ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
5	Awalurahman and Budi (2024)	2024	Conf. Proc.	Paragraph vs sentence in automatic question ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
6	Bahri and Oueslati (2025)	2025	Conf. Proc.	Towards a RAG-based framework for ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
7	Berger et al. (2024)	2024	Conf. Proc.	Question generation capabilities of “small” ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
8	Bernard and Graf (2025)	2025	Conf. Proc.	Language models for educational question ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
9	Bulathwela et al. (2023)	2023	Conf. Proc.	Scalable educational question generation ...	1 ✓	1 ✓	1 ✓	0 ✗	0.5 ~	1 ✓	1 ✓	5.5
10	Contreras-Arquello et al. (2025)	2025	Conf. Proc.	Automatic generation of summaries and ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
11	de-Fitero-Dominguez et al. (2024)	2024	Journal	Automated multiple-choice question ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
12	Deroy et al. (2025)	2025	Journal	Investigating legal question generation ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
13	Deroy et al. (2024)	2024	Preprint	MIRROR: A novel approach for the automated ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	0.5 ~	6

No	Citation	Year	Type	Title (abbreviated)	QA1	QA2	QA3	QA4	QA5	QA6	QA7	Total
14	Elkins et al. (2023)	2023	Conf. Proc.	How useful are educational questions ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
15	J. Feng and Mostow (2021)	2021	Conf. Proc.	Towards difficulty controllable selection of ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
16	W. Feng et al. (2025)	2025	Conf. Proc.	Reasoning and sampling-augmented MCQ ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
17	Fuadi and Wibawa (2022)	2022	Conf. Proc.	Automatic question generation from ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
18	Gala et al. (2021)	2021	Conf. Proc.	Real-time cognitive evaluation of online ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
19	Gomazkova et al. (2024)	2024	Conf. Proc.	Improving efficiency of natural-language ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
20	Grévisse (2024)	2024	Conf. Proc.	Comparative quality analysis of GPT-based ...	1 ✓	1 ✓	1 ✓	0 ✗	1 ✓	1 ✓	1 ✓	6
21	Grévisse et al. (2024)	2024	Journal	Docimological quality analysis of ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
22	Guo et al. (2025)	2025	Conf. Proc.	Research on automatic question generation ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
23	Hwang and Utami (2024)	2024	Journal	Using GPT and authentic contextual ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
24	Jadhav et al. (2025)	2025	Conf. Proc.	Integrated framework for Bloom's based ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
25	Johnson et al. (2025)	2025	Journal	Automatic question generation for Spanish ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
26	Kauchak et al. (2024)	2024	Conf. Proc.	Automatic generation of a large ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
27	Lamsiyah et al. (2024)	2024	Conf. Proc.	Fine-tuning a large language model with...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
28	Lee et al. (2024)	2024	Journal	Few-shot is enough: Exploring ChatGPT...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
29	Levine et al. (2025)	2025	Conf. Proc.	LLM agents for verifiable question ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
30	R. Li et al. (2024)	2024	Conf. Proc.	Generating contextualized mathematics ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6

No	Citation	Year	Type	Title (abbreviated)	QA1	QA2	QA3	QA4	QA5	QA6	QA7	Total
31	Z. Li et al. (2025)	2025	Conf. Proc.	A novel approach to scalable and automatic ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
32	Liao and Koh (2020)	2020	Conf. Proc.	Question generation through transfer ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
33	Lodovico Molina et al. (2024)	2024	Conf. Proc.	Comparison of large language models for ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
34	Ma et al. (2025)	2025	Conf. Proc.	Activate integrated controllable generation ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
35	Maity et al. (2024)	2024	Journal	Investigating large language models for ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
36	Maity et al. (2025a)	2025	Journal	Can large language models meet the ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
37	Maity et al. (2025b)	2025	Preprint	Leveraging in-context learning and ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	0.5 ~	5.5
38	Mary et al. (2025)	2025	Conf. Proc.	Enhancing question answer generation from ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
39	Moore et al. (2023)	2023	Conf. Proc.	Assessing the quality of multiple-choice ...	1 ✓	1 ✓	1 ✓	0 ✗	1 ✓	1 ✓	1 ✓	6
40	Mulla and Gharpure (2023)	2023	Journal	Leveraging well-formedness and cognitive ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
41	Navas et al. (2025)	2025	Conf. Proc.	Exams as a service: Synergies between ...	1 ✓	1 ✓	1 ✓	0 ✗	0.5 ~	1 ✓	1 ✓	5.5
42	D. H. Nguyen et al. (2025)	2025	Conf. Proc.	Enhancing Vietnamese automatic question ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
43	H. A. Nguyen et al. (2022)	2022	Conf. Proc.	Towards generalized methods for automatic ...	1 ✓	1 ✓	1 ✓	0 ✗	0.5 ~	1 ✓	1 ✓	5.5
44	Poonkuzhali et al. (2025)	2025	Conf. Proc.	GenAI-powered adaptive assessments for...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
45	Saha (2021)	2021	Journal	Towards development of a system for ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
46	Sayed et al. (2025)	2025	Conf. Proc.	Subjective question bank generation using ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6

No	Citation	Year	Type	Title (abbreviated)	QA1	QA2	QA3	QA4	QA5	QA6	QA7	Total
47	Scaria et al. (2024)	2024	Conf. Proc.	Automated educational question generation ...	1 ✓	1 ✓	1 ✓	1 ✓	1 ✓	1 ✓	1 ✓	7
48	Shah et al. (2024)	2024	Journal	Improvised progressive model based on ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
49	Shwe et al. (2024)	2024	Journal	The unified difficulty ranking mechanism ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
50	Snegaa et al. (2025)	2025	Conf. Proc.	Advanced learning system—Automatic MCQ ...	1 ✓	1 ✓	1 ✓	0 ✗	1 ✓	1 ✓	1 ✓	6
51	Steuer et al. (2021)	2021	Conf. Proc.	On the linguistic and pedagogical quality ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
52	Suhartono et al. (2024)	2024	Journal	Towards automatic question generation using ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
53	Susanti et al. (2020)	2020	Journal	Integrating automatic question generation ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
54	Timosh et al. (2025)	2025	Journal	Automated question generation with ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
55	Tomikawa and Uto (2024)	2024	Conf. Proc.	Difficulty-controllable multiple-choice ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
56	Van Campenhout et al. (2021)	2021	Conf. Proc.	Transforming textbooks into learning by ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
57	Vincentio and Suhartono (2022)	2022	Journal	Automatic question generation using ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
58	H.-C. Wang et al. (2024)	2024	Journal	A method for generating course test ...	1 ✓	1 ✓	1 ✓	0 ✗	0.5 ~	1 ✓	1 ✓	5.5
59	L. Wang et al. (2025)	2025	Journal	Exploring prompt pattern for generative ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
60	Zhiyun Wang et al. (2025)	2025	Journal	GISedu-GPT: A large language model ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
61	Zichao Wang et al. (2022)	2022	Conf. Proc.	Towards human-like educational question ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
62	Winata et al. (2025)	2025	Conf. Proc.	Utilizing large language models for ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6

No	Citation	Year	Type	Title (abbreviated)	QA1	QA2	QA3	QA4	QA5	QA6	QA7	Total
63	Wiwatbutsiri et al. (2022)	2022	Conf. Proc.	Question generation in the Thai language ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
64	Wróblewska et al. (2025)	2025	Conf. Proc.	Evaluating LLM-generated Q&A test: A ...	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	1 ✓	6.5
65	Wu et al. (2023)	2023	Conf. Proc.	Towards data-effective educational question ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
66	Xu et al. (2024)	2024	Conf. Proc.	Personalized questioning teaching mode for ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
67	L. Yang et al. (2025)	2025	Journal	Tailored definitions with easy reach ...	1 ✓	1 ✓	1 ✓	1 ✓	0.5 ~	1 ✓	1 ✓	6.5
68	Z. Yang et al. (2024)	2024	Journal	Topic-oriented controlled text generation ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
69	Yin and Li (2025)	2025	Journal	KPQDG: A key sentence prompt-tuning for ...	1 ✓	1 ✓	1 ✓	0.5 ~	0.5 ~	1 ✓	1 ✓	6
Summary: 69 included studies (all passed QA1–QA3)					69/69 100%	69/69 100%	69/69 100%	61/69 88.4%	69/69 100%	69/69 100%	69/69 100%	5.5– 7.0

Notes: QA1 = Primary empirical research; QA2 = Focus on AQG; QA3 = Uses Transformer/LLM; QA4 = Cognitive control methods; QA5 = Quality validation beyond automated metrics; QA6 = Describes corpus/dataset; QA7 = Presents findings clearly. Summary counts report the studies that met each criterion at a score of 0.5 or above, consistent with Figure 6; a full score of 1.0 was recorded for 10 studies on QA4, 17 on QA5, and 67 on QA7. Indonesian-language studies (bold): Atmajaya et al. (2025), Awalurahman and Budi (2024), Fuadi and Wibawa (2022), Suhartono et al. (2024), and Vincentio and Suhartono (2022).

AUTHORS



Firmansyah Ibrahim is a PhD researcher in the Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia, and a lecturer in the Department of Information Systems, Universitas Islam Negeri Alauddin Makassar, Indonesia. His research advances natural language processing for resource-constrained languages by developing a controlled automatic question-generation system. He can be reached at firmansyah.ibrahim.2505349@students.um.ac.id or firmansyah.ibrahim@uin-alauddin.ac.id. (<https://orcid.org/0009-0007-9071-059X>)



Dr Eng. Didik Dwi Prasetya is a Senior Lecturer in the Department of Electrical Engineering and Informatics at Universitas Negeri Malang, Indonesia. His research focuses on information technology, with specific interests in AI in education, educational data science, natural language processing, knowledge engineering, and multiplatform software engineering. (<https://orcid.org/0000-0002-3540-2961>)



Dr Ir. Triyanna Widiyaningtyas, M.T. is a lecturer and researcher in the Department of Electrical Engineering and Informatics at Universitas Negeri Malang, Indonesia. Her research primarily focuses on software engineering, data mining, database systems, and recommender systems. (<https://orcid.org/0000-0001-6104-6692>)