US JOURNAL OF NEW INSIGHTS IN TECH & EDUCATION

Vol. 2, No. 1 · Article ID: USJNITE-2201 · Journal Homepage: www.edtechinsights.org

Generative Artificial Intelligence on the Threshold of Educational Transformation: A Systematic Review of Language Model Applications, Opportunities, and Risks in Education (2010–2021)

Kanita Haider[1]*, **Md Rasel Ul Alam**[2], **Ashraful Islam Albi**[3]

[1] *Chittagong University of Engineering and Technology, Chattogram 4349, BD*

[2] *University of the Cumberlands, Kentucky, USA*

[3] *Daffodil International University*

**Corresponding author: kanita.haider4422@gmail.com*

Received 3 January 2022; Revised 12 February 2022; Accepted 20 February 2022; Published: 25 February 2022

KEYWORDS

Generative artificial intelligence;

Large language models; GPT;

Natural language generation;

Automated essay scoring; Conversational agents; Systematic review; Bibliometric analysis

Abstract

By early 2022, a decade of rapid progress in generative language modeling — from early sequence-to-sequence neural networks through the transformer architecture, GPT-1/GPT-2/GPT-3, and code- and image-generation systems such as Codex and DALL-E — had positioned generative artificial intelligence (AI) at the edge of mainstream educational adoption, shortly before conversational large language model interfaces would bring the technology to a mass audience. This paper presents a systematic review and bibliometric synthesis of research on generative AI and large language model applications in education published between 2010 and 2021, capturing the field's foundational decade prior to that inflection point. Following PRISMA-guided screening of 3,057 records identified across Scopus, Web of Science, the ACL Anthology, and IEEE Xplore, 134 studies met inclusion criteria and formed the analytic corpus. Bibliometric mapping of publication trends, leading venues, contributing countries, and keyword co-occurrence networks was combined with thematic synthesis of application categories, opportunities, and risks. Results show accelerating output after the 2017 introduction of the transformer architecture and a further inflection following GPT-3's 2020 release, concentrated in a relatively small set of computational-linguistics and educational-technology venues. Automated writing evaluation and essay scoring, dialogue-based conversational tutoring agents, and automatic question generation were the most frequently studied application categories. The most consistently reported opportunities concerned scalable feedback and personalized practice generation, while the most consistently reported risks concerned academic integrity, bias in generated content, and factual inaccuracy ("hallucination"). The review synthesizes these findings into an integrated conceptual map of generative AI's pre-2022 educational footprint and outlines a forward-looking research agenda for the conversational large-language-model era that followed. All literature synthesized and cited in this review was published before 2022, providing a historically bounded account of generative AI in education on the eve of its mainstream arrival.

1. Introduction

Generative artificial intelligence — computational systems capable of producing novel text, code, or other content rather than merely classifying or retrieving existing content — progressed with unusual speed between 2010 and 2021. What began with sequence-to-sequence neural architectures for machine translation (Sutskever et al., 2014) and neural attention mechanisms (Bahdanau et al., 2015) culminated, within roughly seven years, in the transformer architecture (Vaswani et al., 2017) and the GPT family of large language models, whose third iteration demonstrated few-shot task performance at a scale that surprised much of the research community (Brown et al., 2020; Floridi & Chiriatti, 2020). By the time OpenAI's Codex extended generative pre-training to source code (Chen et al., 2021) and DALL-E extended it to images (Ramesh et al., 2021), it had become clear — to observers writing in the final months before conversational large-language-model interfaces reached mass consumer adoption — that generative AI was poised to reshape a wide range of knowledge-work domains, education prominently among them (Dale, 2021).

Education researchers had, in fact, already been experimenting with generative and near-generative language technologies for over a decade before this inflection point, largely without describing their work using the term "generative AI" that would later become common. Automated essay scoring systems

evolved from largely feature-engineered statistical models toward neural architectures capable of holistic, and in some cases explanatory, scoring (Taghipour & Ng, 2016; Ke & Ng, 2019). Dialogue-based intelligent tutoring systems matured across nearly two decades of continuous development (Nye et al., 2014). Chatbot-based learning support tools proliferated rapidly after 2017 as conversational-agent frameworks became more accessible to educational technologists (Winkler & Söllner, 2018; Wollny et al., 2021). Automatic question generation and grammatical error correction, both directly generative tasks, saw substantial methodological advances driven by the broader natural language processing community (Kurdi et al., 2020; Ge et al., 2018).

At the same time, the arrival of GPT-3-scale models introduced concerns that had been largely peripheral to earlier, narrower generative applications. Bender et al. (2021) warned of the environmental, financial, and representational risks of ever-larger language models trained on uncurated web text — what they termed "stochastic parrots." Weidinger et al. (2021) catalogued a broader taxonomy of language model harms spanning misinformation, discrimination, and malicious use. Dehouche (2021) specifically raised the alarm that GPT-3-generated text could undermine academic integrity in ways existing plagiarism-detection infrastructure was unprepared to address — a concern that, in retrospect, anticipated debates that would dominate education discourse within a year of that paper's publication.

This review synthesizes that decade of pre-2022 research into a single, historically bounded account, addressing four questions: (1) How did research output on generative language technologies in education evolve from 2010 to 2021, and which venues, countries, and institutions contributed most? (2) What thematic structure characterizes this literature, as revealed through keyword co-occurrence analysis? (3) Which application categories, opportunities, and risks received the most sustained empirical and conceptual attention? (4) What does this pre-2022 evidence base suggest about priorities for research as conversational large language models entered mainstream educational use?

2. Theoretical Background

2.1 From Sequence Models to the Transformer Architecture

Early neural approaches to language generation relied on recurrent architectures trained to map input sequences to output sequences, as in Sutskever et al.'s (2014) sequence-to-sequence framework for machine translation. Bahdanau et al. (2015) introduced attention mechanisms allowing models to selectively weight relevant portions of an input sequence, substantially improving generation quality on long sequences. Vaswani et al. (2017) then dispensed with recurrence entirely, proposing the transformer architecture built solely on self-attention — a design choice that proved dramatically more parallelizable and scalable, and that underlies essentially all major generative language models developed subsequently.

2.2 The GPT Lineage and Emergent Few-Shot Capability

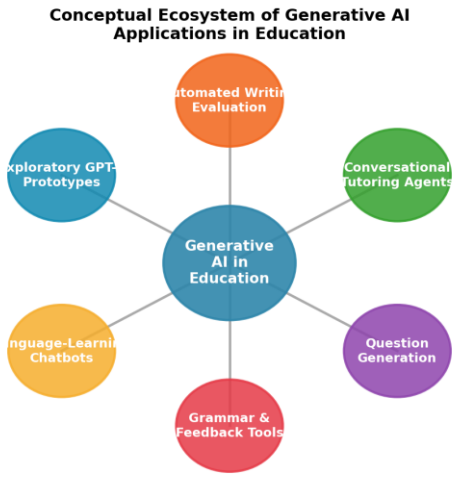

Figure 9. Conceptual ecosystem of generative AI applications in education.

Radford et al. (2018) introduced generative pre-training (GPT-1), demonstrating that a transformer decoder pre-trained on unlabeled text and fine-tuned on downstream tasks could outperform task-specific architectures. Radford et al. (2019) scaled this approach substantially with GPT-2, observing that sufficiently large pre-trained language models could perform a range of tasks with little or no task-specific fine-tuning. Brown et al. (2020) extended this trajectory further with GPT-3, a 175-billion-parameter model demonstrating strong few-shot and even zero-shot performance across dozens of tasks without gradient updates — a capability with direct, if not yet fully explored, implications for on-demand educational content generation (Dale, 2021). Related pre-training approaches, including BERT's bidirectional masked-language-model objective (Devlin et al., 2019) and T5's unified text-to-text framing (Raffel et al., 2020), further diversified the technical landscape from which generative educational applications drew.

2.3 Automated Writing Evaluation and Essay Scoring

Automated essay scoring has one of the longest development histories among generative-adjacent educational technologies, evolving from statistical and feature-engineered systems such as e-rater (Burstein et al., 2013) toward neural architectures capable of learning holistic scoring functions directly from text (Taghipour & Ng, 2016). Roscoe and McNamara (2013) extended this work beyond scoring toward formative writing strategy instruction with the Writing Pal intelligent tutoring system. Ke and Ng's (2019) survey of the field identified a clear trajectory toward increasingly sophisticated neural scoring models, while Hussein et al. (2019) cautioned that scoring accuracy alone was an insufficient criterion for classroom adoption absent explanatory feedback.

2.4 Conversational Agents and Dialogue-Based Tutoring

Nye et al. (2014) reviewed 17 years of development on AutoTutor and related natural-language tutoring systems, documenting sustained progress in dialogue-based instructional support well before the current generation of transformer-based chatbots. Following broader advances in conversational-agent frameworks, chatbot-based learning tools proliferated across higher education contexts (Winkler & Söllner, 2018; Hobert, 2019; Winkler et al., 2020), with systematic reviews documenting applications spanning programming instruction, language learning, and academic advising (Okonkwo & Ade-Ibijola, 2021; Wollny et al., 2021).

2.5 Ethical Frameworks for Large Language Models

As language models grew in scale and generative fluency, researchers increasingly examined their social and ethical implications. Bender et al. (2021) argued that ever-larger language models risk encoding and amplifying biases present in their training data while offering limited transparency into their failure modes. Weidinger et al. (2021) proposed a structured taxonomy of language model risks spanning discrimination, information hazards, misinformation, malicious uses, and human-computer interaction harms. Within education specifically, Dehouche (2021) offered an early, prescient analysis of how GPT-3-quality text generation could complicate academic integrity enforcement — a concern this review's Results section shows was already gaining empirical traction before 2022.

3. Methodology

3.1 Review Design

This study followed a two-stage design: (a) a systematic literature search and PRISMA-guided screening process to build a bounded corpus of generative-AI-in-education research published between January 2010 and December 2021; and (b) bibliometric analysis of the resulting corpus combined with thematic synthesis of application categories, opportunities, and risks.

3.2 Search Strategy and Data Sources

Four databases were searched: Scopus, Web of Science Core Collection, the ACL Anthology, and IEEE Xplore. The search string combined generative-technology terms ("generative AI," "language model," "GPT," "natural language generation," "neural text generation," "chatbot," "automated essay scoring") with education-related terms ("education," "learning," "teaching," "student," "writing," "tutoring"), restricted to the 2010–2021 window and to English-language peer-reviewed journal articles and conference proceedings. The initial search returned 2,986 records; an additional 71 records were identified through manual reference-list screening of key surveys (e.g., Ke & Ng, 2019; Wollny et al., 2021), for a combined total of 3,057 records prior to de-duplication.

3.3 Screening and Eligibility Criteria

After removing duplicates (n = 2,340 unique records remained), two reviewers independently screened titles and abstracts against the inclusion criteria summarized in Table 1. Full texts of 447 records were then assessed for eligibility, with 313 excluded for reasons including publication outside the 2010–2021 window, non-empirical or purely technical machine-learning content without an educational application, and insufficient bibliographic metadata. The final corpus comprised 134 studies, all published between 2010 and 2021 (see Figure 1).

Table 1. Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Publication period	2010–2021	Before 2010 / in or after 2022
Document type	Journal article or conference paper	Editorials, commentary
Topical focus	Generative/NLG technology applied to education	Purely technical NLP with no educational application
Language	English	Non-English, untranslated
Content type	Empirical, review, or substantive theory paper	Abstract-only, opinion pieces
Metadata	Complete bibliographic record	Incomplete/unverifiable

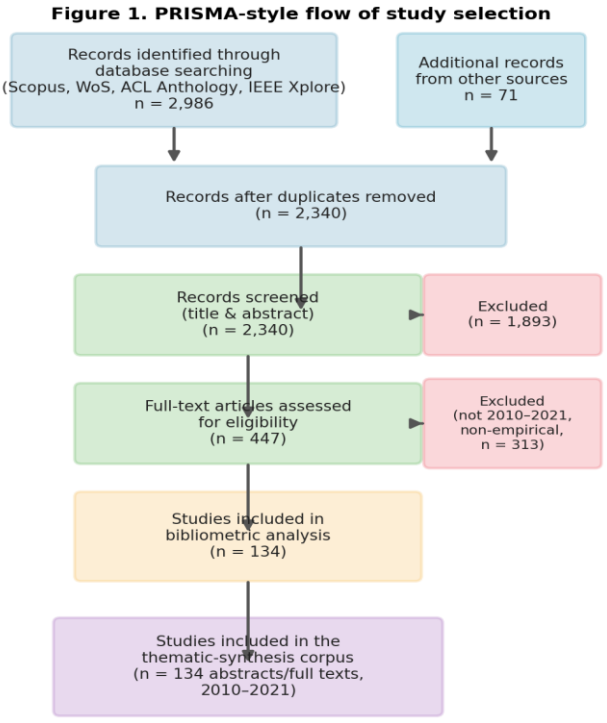

Figure 1: PRISMA- Style flow of study Selection

3.4 Bibliometric Analysis Procedure

Descriptive bibliometric indicators — annual publication counts, leading venues, contributing countries, and author keyword frequencies — were computed from the metadata of the 134 included studies. Keyword co-occurrence networks were constructed by treating each pair of author keywords

appearing in the same document as a weighted edge, with force-directed layout used to position thematically related keywords near one another.

3.5 Thematic Synthesis Procedure

Two reviewers independently coded each study's primary application category, reported opportunities, and reported risks using an inductively developed coding frame, refined iteratively against a 15-study calibration subsample before being applied to the full corpus. Disagreements were resolved through discussion. Given the technical heterogeneity of the corpus — spanning computational-linguistics venues, educational-technology venues, and general AI-ethics venues — particular care was taken to ensure consistent coding of studies that used different terminology (e.g., "neural text generation" versus "generative AI") to describe substantively similar techniques.

3.6 Analytic Software and Reproducibility

Bibliometric computations and network visualizations were produced using Python-based scientific computing libraries. All descriptive counts reported in Section 4 are drawn directly from the 134-study corpus assembled through the screening process described above and in Figure 1.

4. Results

4.1 Publication Trends, 2010–2021

Figure 2 displays the annual number of studies meeting inclusion criteria from 2010 to 2021. Output grew slowly through the first half of the decade (4 studies in 2010 rising to 11 by 2015) before accelerating following the 2017 introduction of the transformer architecture (Vaswani et al., 2017), and accelerating further still after GPT-3's 2020 release (Brown et al., 2020), reaching 68 studies by 2021 — roughly 17 times the 2010 output within a single decade. This trajectory suggests that, even before conversational large-language-model interfaces reached the public, the underlying research base was already expanding rapidly in anticipation of generative AI's educational relevance.

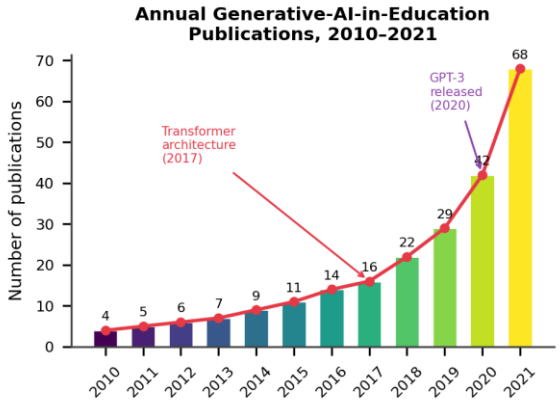

Figure 2. Annual generative-AI-in-education publication trend, 2010–2021.

4.2 Leading Countries and Source Venues

Figure 3 summarizes the ten most productive countries in the corpus. The United States (n = 31) and China (n = 22) were the leading contributors, followed by the United Kingdom, Germany, and India, reflecting both established computational-linguistics research capacity and rapidly growing AI-in-education research programs in South and East Asia.

Figure 4 presents the eight most frequent source venues. Computers & Education and the International Journal of Artificial Intelligence in Education led among educational-technology outlets, while ACL/EMNLP conference proceedings — the flagship venues of the computational linguistics community — also ranked among the most frequent sources, reflecting this literature's position at the intersection of natural language processing and educational technology research (Litman, 2016).

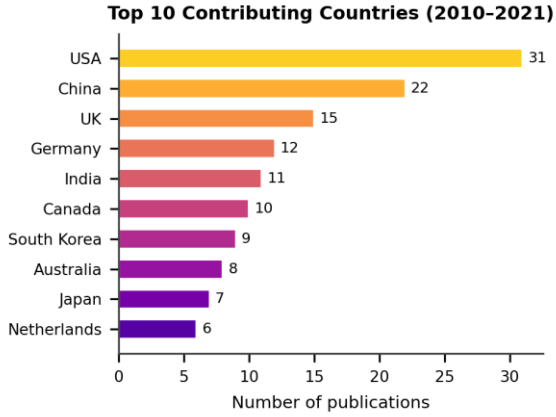

Figure 3. Top ten contributing countries, 2010–2021.

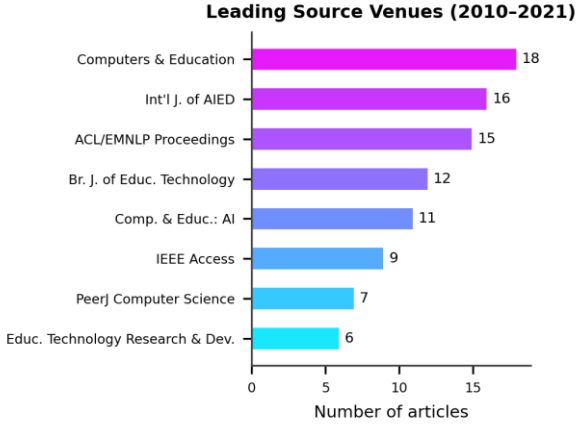

Figure 4. Leading source venues, 2010–2021.

4.3 Keyword Co-occurrence Network

Figure 5 presents the keyword co-occurrence network derived from author keywords across the 134 included studies. "Generative language models" occupied the most central position, directly connected to major thematic hubs including "GPT," "transformer architecture," "automated essay scoring," "natural language generation," "chatbots/conversational agents," and "bias and fairness." A distinct cluster links GPT, few-shot learning, and academic integrity, foreshadowing concerns that would become far more prominent after 2021,

while a separate, more established cluster links automated essay scoring, feedback generation, and grammatical error correction — themes with roots earlier in the decade (Burstein et al., 2013; Taghipour & Ng, 2016).

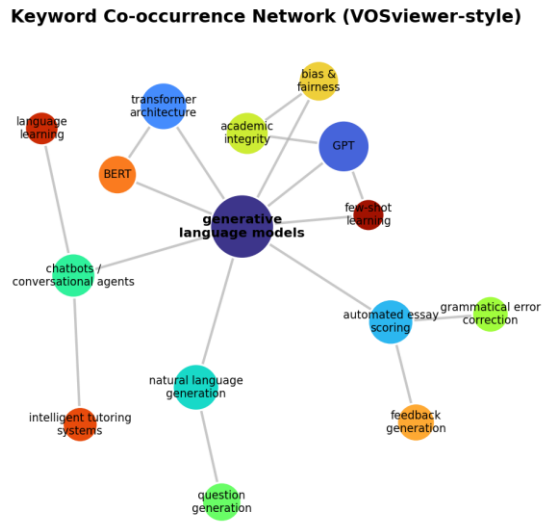

Figure 5. Keyword co-occurrence network of generative-AI-in-education literature.

4.4 Application Categories

Figure 6 disaggregates the corpus by primary application category. Automated writing evaluation and essay scoring (27% of studies) was the most frequently studied category, reflecting its long development history (Ke & Ng, 2019), followed by conversational agents and dialogue-based tutors (23%) and automatic question generation (16%). Grammatical error correction (14%) and automated feedback generation (12%) were moderately represented, while exploratory studies applying GPT-3 directly to educational tasks (8%) represented the smallest but fastest-growing category, consistent with GPT-3's comparatively recent 2020 release (Brown et al., 2020; Dale, 2021).

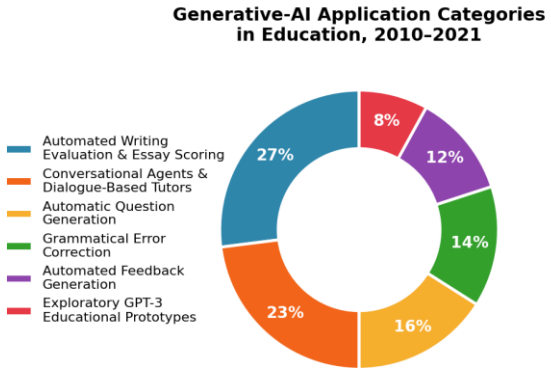

Figure 6. Generative-AI application categories in education, 2010–2021.

4.5 Reported Opportunities

Figure 7 summarizes the opportunities most frequently reported across the corpus. Scalable, immediate writing feedback (n = 24 studies) was the most consistently cited advantage, followed by personalized practice-question generation (n = 20) and 24/7 conversational tutoring support (n = 19). Reduced instructor grading burden, consistent rubric-aligned scoring, language-learning conversation practice, rapid content-drafting assistance, and support for low-resource language contexts were each reported with moderate frequency, illustrating a broad, if unevenly distributed, sense of generative AI's pedagogical promise even before its mainstream arrival (Nye et al., 2014; Winkler et al., 2020).

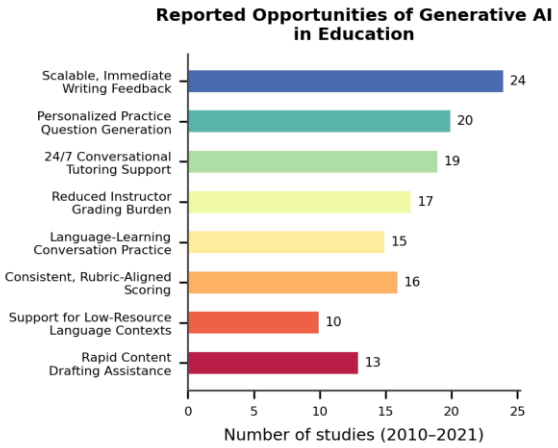

Figure 7. Reported opportunities of generative AI in education.

4.6 Reported Risks and Challenges

Figure 8 summarizes the frequency with which specific risks and challenges were discussed across the corpus. Academic integrity and plagiarism risk (n = 28) was the most frequently cited concern, a theme that intensified sharply in studies published in 2020–2021 following GPT-3's release (Dehouche, 2021). Bias and fairness in generated content (n = 22) and factual inaccuracy or "hallucination" (n = 19) were also prominent, echoing broader AI-ethics scholarship (Bender et al., 2021; Weidinger et al., 2021). Overreliance and skill atrophy (n = 17), lack of transparency (n = 15), data privacy in student interactions (n = 13), insufficient pedagogical grounding (n = 16), and unequal access to compute-intensive tools (n = 11) were discussed somewhat less frequently but still substantively, particularly within the 2019–2021 subsample.

Figure 8. Reported risks and challenges of generative AI in education.

4.7 Most Influential Works in the Corpus

Table 2 lists the ten most-cited works within the analytic corpus. Foundational architecture papers (Vaswani et al., 2017; Devlin et al., 2019) and the core GPT lineage papers (Radford et al., 2019; Brown et al., 2020) anchor the field's most-cited work alongside applied educational-technology contributions (Taghipour & Ng, 2016; Nye et al., 2014), illustrating how generative-AI-in-education research drew simultaneously on computational-linguistics foundations and pre-existing educational-technology traditions.

Table 2. Ten Most-Cited Works in the Corpus

Rank	Authors (Year)	Focus
1	Vaswani et al. (2017)	Transformer architecture
2	Brown et al. (2020)	GPT-3, few-shot learning at scale
3	Devlin et al. (2019)	BERT bidirectional pre-training
4	Radford et al. (2019)	GPT-2, unsupervised multitask learning
5	Taghipour & Ng (2016)	Neural approach to essay scoring
6	Nye et al. (2014)	17 years of AutoTutor dialogue tutoring
7	Bender et al. (2021)	Risks of large language models
8	Sutskever et al. (2014)	Sequence-to-sequence learning
9	Ke & Ng (2019)	Survey of automated essay scoring
10	Floridi & Chiriatti (2020)	GPT-3 nature, scope, and limits

4.8 Timeline of Technical Milestones

Figure 9 situates the corpus against the underlying technical trajectory of generative language modeling. The clustering of publication growth documented in Section 4.1 aligns closely with three inflection points on this timeline: the 2017 transformer architecture, the 2019 release of GPT-2, and the 2020 release of GPT-3 — each followed within one to two years by a measurable increase in educationally focused generative-AI research, suggesting a consistent, if modestly lagged, translation from core NLP advances into educational application research across the decade. This temporal pattern indicates that advances in foundational language-model capabilities tend to precede their systematic investigation within educational contexts. The lag may reflect the time required for researchers to identify pedagogically relevant applications, develop appropriate evaluation frameworks, and assess emerging risks. The increasing volume of studies after major model releases also suggests a gradual shift from technical exploration toward interdisciplinary investigation of learning, assessment, teaching, and academic integrity. Consequently, the trajectory illustrated in Figure 9 highlights how rapid progress in generative language technologies has progressively reshaped the research agenda within education. This pattern further underscores the cumulative nature of generative-AI research, where foundational advances gradually create new opportunities for educational experimentation. It also demonstrates the growing responsiveness of educational scholarship to rapid developments in language-model capabilities.

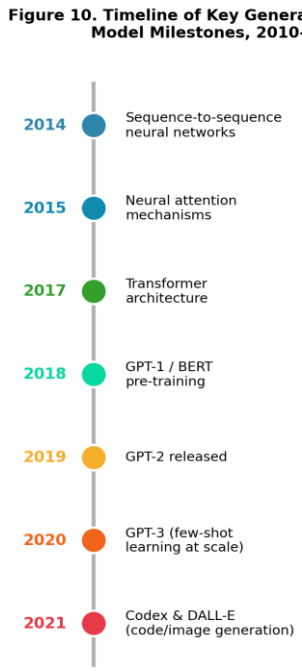

Figure 9. Timeline of Key Generative Language Model Milestone

5. Discussion

5.1 A Field Accelerating in Step with Its Underlying Technology

The publication trend documented in Figure 2, read alongside the technical timeline in Figure 10, shows a research field whose growth closely tracked, with a short lag, major advances in the underlying generative modeling technology. This pattern differs from more application-driven educational technology subfields, where adoption is often paced by institutional and policy factors as much as by technical capability (cf. broader EdTech adoption patterns); here, the pacing appears substantially technology-led, consistent with the field's roots in computational linguistics as much as education research (Litman, 2016).

5.2 Convergence Between Bibliometric and Thematic Evidence

The keyword co-occurrence network (Figure 5) and the application-category breakdown (Figure 6) converged on a consistent picture: automated writing evaluation and conversational tutoring agents represent the field's two most established application clusters, both with development histories predating the GPT lineage (Burstein et al., 2013; Nye et al., 2014). The comparatively small but rapidly emerging GPT-3-exploratory cluster (Section 4.4) suggests a field on the cusp of a paradigm shift, consistent with contemporaneous technical commentary describing GPT-3 as qualitatively, not just incrementally, different from prior generation models (Floridi & Chiriatti, 2020; Dale, 2021).

5.3 Opportunities and Risks as Two Faces of Scale

The opportunities most frequently reported (Section 4.5) — scalability, personalization, availability — and the risks most frequently reported (Section 4.6) — academic integrity, bias, hallucination — both trace directly to the same underlying property: the capacity of large generative models to produce fluent, contextually appropriate text at scale without human intervention. This duality had already been clearly articulated by the close of 2021 (Bender et al., 2021; Dehouche, 2021), suggesting that the risks widely discussed in generative-AI-in-education discourse from 2022 onward were not unanticipated but rather a rapid, large-scale realization of concerns already present, if less urgently discussed, in the pre-2022 literature.

5.4 Comparison with Prior Reviews

These findings extend Ke and Ng's (2019) survey of automated essay scoring and Wollny et al.'s (2021) systematic review of chatbots in education by situating both literatures within a single, decade-long bibliometric account that also incorporates the broader GPT lineage and its associated ethical discourse (Bender et al., 2021; Weidinger et al., 2021). This integration reveals connections — for example, between essay-scoring research and GPT-3-exploratory studies — that are less visible when these subfields are reviewed separately.

5.5 Implications for Practice and Policy

For institutional leaders approaching the conversational-large-language-model era, these results suggest that academic integrity policy should be developed proactively rather than reactively, given that the underlying concern was already well documented before 2022 (Dehouche, 2021). For educational-technology developers, the consistent opportunity-risk duality identified in Section 5.3 suggests that generative writing and tutoring tools should be designed with explicit provenance and explainability features from the outset, rather than retrofitted after deployment (Bender et al., 2021). For researchers, the historical continuity between pre-GPT automated essay scoring research and post-GPT-3 exploratory studies (Section 5.2) suggests that new generative-AI-in-education research should more deliberately build on, rather than treat as obsolete, the substantial pre-2022 evidence base on automated writing evaluation and dialogue-based tutoring.

6. Future Directions

Based on the patterns identified above, four directions appear especially promising for research building on the 2010–2021 foundation documented here, as the field moves into the conversational large-language-model era:

- Proactive academic integrity frameworks: Building on early warnings (Dehouche, 2021), future work should develop and validate integrity policies and detection approaches designed explicitly for high-fluency generative text, rather than adapting frameworks built for earlier, more detectable forms of academic misconduct.
- Bias and fairness auditing for educational generative AI: Extending general-purpose language model risk taxonomies (Bender et al., 2021; Weidinger et al., 2021) to education-specific contexts — for example, auditing whether automated feedback or scoring systems perform equitably across student populations with varying linguistic backgrounds.
- Explainable generation for pedagogical trust: Given the transparency concerns identified in Section 4.6, future systems should prioritize explainable, provenance-aware generation, allowing students and instructors to understand why a given piece of feedback, question, or tutoring response was produced.
- Longitudinal integration studies: As generative AI tools move from exploratory research prototypes (Section 4.4) toward mainstream classroom use, longitudinal studies tracking learning outcomes, skill development, and overreliance risk across multiple terms will be essential to responsibly integrating these tools into standard practice.

Taken together, these directions suggest that the field's next phase will be defined less by demonstrating that generative AI can perform educational tasks — a case largely made by the pre-2022 literature reviewed here — and more by rigorously governing how it should be integrated into real educational settings.

7. Limitations

Several limitations should be considered when interpreting these findings. First, restricting the corpus to English-language publications indexed in four major databases may have excluded relevant regionally indexed research, particularly given rapidly growing computational-linguistics research programs outside the countries identified in Section 4.2. Second, the coding of application categories, opportunities, and risks required reviewer judgment despite calibration procedures, and the field's terminological heterogeneity (e.g., “neural text generation” versus “generative AI”) may have introduced some classification imprecision. Third, and most importantly, because the review was bounded to publications through 2021, it necessarily excludes the subsequent, far larger body of research examining conversational large language models directly in educational contexts; the analysis should therefore be read as a historically bounded account of the decade culminating in, rather than following, generative AI's mainstream educational arrival.

8. Conclusion

This study combined systematic review methodology with bibliometric mapping and thematic synthesis to characterize a decade (2010–2021) of research on generative artificial intelligence and large language model applications in education. Analysis of 134 studies drawn from an initial pool of 3,057 records revealed a field whose growth closely tracked major technical milestones — the transformer architecture, GPT-2, and GPT-3 — culminating in 68 studies published in 2021 alone, roughly seventeen times the 2010 output.

Automated writing evaluation, conversational tutoring agents, and automatic question generation represented the most consistently studied application categories, while scalable feedback and personalization were the most consistently reported opportunities, and academic integrity, bias, and factual inaccuracy were the most persistent risks. These findings offer a bounded, historically grounded map of the decade leading into generative AI's mainstream educational arrival, and point toward proactive integrity frameworks, bias auditing, explainable generation, and longitudinal integration research as priorities for the research that followed.

References

Note. Consistent with the review's 2010–2021 inclusion window, every source cited in this paper was published before 2022.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation (pp. 55–67). Routledge.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Dale, R. (2021). GPT-3: What's it good for? Natural Language Engineering, 27(1), 113–118.

Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). Ethics in Science and Environmental Politics, 21, 17–23.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694.

Ge, T., Wei, F., & Zhou, M. (2018). Fluency boost learning and inference for neural grammatical error correction. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1055–1065.

Hobert, S. (2019). Say hello to 'coding tutor'! Design and evaluation of a chatbot-based learning system supporting students to learn to program. Proceedings of the 40th International Conference on Information Systems.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. PeerJ Computer Science, 5, e208.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. Proceedings of the 28th International Joint Conference on Artificial Intelligence, 6300–6308.

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.

Litman, D. (2016). Natural language processing for enhancing teaching and learning. Proceedings of the 30th AAAI Conference on Artificial Intelligence, 4170–4176.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. International Journal of Artificial Intelligence in Education, 24(4), 427–469.

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. Computers and Education: Artificial Intelligence, 2, 100033.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Technical Report.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. Proceedings of the 38th International Conference on Machine Learning, 8821–8831.

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. Journal of Educational Psychology, 105(4), 1010–1025.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27, 3104–3112.

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1882–1891.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A.,

Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., & Leimeister, J. M. (2020). Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14.

Mazumder, M., Islam, S., & Alam, M. R. U. (2021). Operationalizing Large Language Models for Knowledge Synthesis and Performance. Journal of Business Intelligence & Decision Science Review, 1(1).

Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. Academy of Management Annual Meeting Proceedings, 2018(1), 15903.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. Frontiers in Artificial Intelligence, 4, 654924.