sustainability

MDPI

Systematic Review

Environmental Impacts of Generative AI in Education: A Systematic Review of Educational and Technical Evidence

Marko Radovan *, Tadej Košmerl and Danijela Makovec Radovan

Faculty of Arts, University of Ljubljana, 1000 Ljubljana, Slovenia; tadej.kosmerl@ff.uni-lj.si (T.K.); danijela.makovec@ff.uni-lj.si (D.M.R.)
* Correspondence: marko.radovan@ff.uni-lj.si

Abstract

The rapid adoption of generative artificial intelligence (GenAI) in educational institutions has occurred without systematic assessment of its environmental impacts. Training and operating large language models result in electricity consumption, greenhouse gas emissions, water use, hardware manufacturing, and electronic waste, yet these physical footprints remain largely invisible to educators and learners. This PRISMA-based systematic review addressed two research questions: (a) What types of environmental impacts are associated with GenAI tools in educational contexts? and (b) What indicators, data sources, assumptions, and methodological approaches are used to measure or estimate these impacts? A search of four databases (Scopus, Web of Science, ERIC, IEEE Xplore) identified 23 eligible publications from 2022 to 2026, classified into two evidence streams: education-specific studies (n = 8) examined GenAI in direct educational settings, while complementary technical studies (n = 15) provided transferable environmental indicators, benchmarks, and assessment methods from the broader AI sustainability literature. This two-stream design was chosen because education-specific environmental evidence remains scarce, and methodological approaches from technical literature are essential for understanding how educational impacts could be assessed. The synthesis found that operational electricity consumption, carbon emissions, and water demand are the most frequently reported impacts, but evidence in educational settings remains limited and methodologically inconsistent. Technical studies provide precise hardware-level measurements and lifecycle assessments, while educational research relies mainly on indirect proxies, self-reported surveys, or token-count approximations. Education-specific empirical evidence remains limited, with a small number of studies suggesting that GenAI use can increase the energy footprint of student work under some conditions, while real-time carbon-feedback displays may reduce prompt volume and estimated emissions. Comparative labor studies provide mixed findings: although some estimates portray AI-generated work as less carbon-intensive than human labor, correctness-controlled evidence from programming tasks shows that emissions can increase substantially when iterative prompting and output verification are included. Water footprints, embodied emissions, and electronic waste are acknowledged conceptually but measured empirically only in related technical fields. The review proposes a three-tiered policy framework spanning pedagogical practices, technical platform choices, and institutional governance to align educational AI use with environmental sustainability commitments.

Keywords: generative artificial intelligence; educational technology; environmental impact; systematic review; sustainability metrics

Check for updates

Received: 23 June 2026
Revised: 8 July 2026
Accepted: 10 July 2026
Published: 15 July 2026

1. Introduction

Generative artificial intelligence has rapidly entered educational practice. Students use large language models and image generators for writing, coding, translation, feedback, and creative work, while teachers increasingly apply these systems to lesson planning, assessment design, administrative tasks, and student support [1–3]. Educational institutions are also beginning to integrate generative AI into learning platforms and institutional services. Most research has focused on pedagogical opportunities, academic integrity, assessment, privacy, and bias, while the environmental consequences of this expansion have received far less systematic attention [4–6].

Generative AI is often experienced as an immaterial service. A user enters a prompt through a browser or application and receives an answer within seconds. Yet this interaction depends on physical infrastructure, including data centers, high-performance processors, cooling systems, electricity grids, network equipment, and hardware supply chains. Training and operating large AI models create environmental impacts through electricity consumption, greenhouse gas emissions, water use, hardware manufacturing, mineral extraction, and electronic waste [7–9]. These impacts extend across the lifecycle of AI systems and cannot be reduced to the electricity consumed during a single training run or user query. Electricity demand has become a particularly visible concern. Data centers consumed an estimated 415 TWh of electricity worldwide in 2024, equivalent to approximately 1.5% of global electricity use, and their electricity consumption has grown considerably faster than total global demand in recent years [10]. AI is not solely responsible for this growth, but the increasing use of computationally intensive training and inference workloads is expected to become an important driver of future data center demand [10].

At the model level, energy consumption varies by model architecture, parameter count, hardware type, workload configuration, batch size, prompt length, output length, and system utilization [11–13]. Carbon emissions depend not only on electricity consumption but also on when and where computation occurs. The same model and task can produce different emissions when hosted in regions with different electricity-generation mixes and carbon intensities [13,14]. Data-center efficiency is also important. Measures such as power usage effectiveness account for the additional electricity required for cooling and other supporting infrastructure, so processor-level measurements alone do not capture the full operational footprint of AI services. Estimates that omit server efficiency, regional energy conditions, and workload characteristics may therefore produce misleading comparisons between models or use cases. Water use adds another layer of environmental concern. AI-related water consumption arises directly through data-center cooling and indirectly through electricity generation. Both sources vary geographically and temporally, as local climate conditions, cooling technologies, and electricity sources affect water efficiency [8]. Research has therefore called for water footprints to be considered alongside energy and carbon metrics rather than treated as a secondary issue [8,15]. This is especially relevant where data centers operate in regions already experiencing water stress. The environmental footprint also begins before a model is trained or deployed. Producing graphics processing units, servers, storage systems, and network equipment requires metals, minerals, chemicals, water, and energy. Manufacturing and transporting this equipment generate embodied emissions, while short hardware replacement cycles contribute to electronic waste and additional material demand [9,16,17]. Lifecycle studies therefore distinguish operational impacts from embodied and end-of-life impacts. Without this broader boundary, assessments risk focusing on visible electricity consumption while overlooking resource extraction, manufacturing, transport, and disposal.

These measurement challenges are especially relevant in education. Educational use is often distributed across large populations and repeated over time. A single student interaction may require relatively little energy, but routine use across courses, institutions, and national education systems can create substantial aggregate demand. Environmental impacts may also vary across educational activities. Generating images, debugging code through repeated exchanges, producing automated feedback for an entire class, or operating an institution-wide chatbot involve different models, workloads, and interaction patterns [18–20]. For this reason, generic estimates of emissions per prompt cannot be transferred uncritically from one educational activity to another.

Education-specific research has only begun to address these issues. Some studies estimate emissions from token counts or model-specific benchmarks, while others examine changes in student behavior after receiving environmental feedback [18,19]. Classroom-based research has also explored changes in total energy use when generative AI is introduced into educational activities [20]. Yet such studies remain uncommon, and many educational publications discuss environmental sustainability only conceptually. They often omit information needed for reproducible assessment, including the model used, number of interactions, input and output tokens, server assumptions, hardware type, system boundaries, power usage effectiveness, grid carbon intensity, and treatment of embodied impacts.

The evidence base is also divided across disciplinary boundaries. Educational studies tend to examine student behavior, awareness, teaching practices, or institutional adoption. Computer science research more often measures processor power draw, inference efficiency, model size, and workload characteristics. Environmental and lifecycle studies contribute methods for estimating carbon, water, manufacturing, and material impacts [9,11,21]. These areas address related questions, but they use different units, system boundaries, assumptions, and evidential standards. As a result, their findings remain difficult to compare or apply directly to educational decision-making. This fragmentation creates both an empirical and a methodological gap. It remains unclear which environmental impacts have been documented specifically in educational contexts, how these impacts have been measured or estimated, and which methods from the broader AI sustainability literature can support more credible educational assessment. Without such synthesis, institutions may treat GenAI services as environmentally weightless, while researchers may rely on simplified proxies that overlook differences in hardware efficiency, data center operations, electricity sources, cooling systems, and hardware lifecycles.

This systematic review examines how the environmental impacts of generative AI are identified, measured, estimated, and reported in relation to educational use. It first maps evidence from education-specific publications, then examines complementary studies from the broader AI sustainability literature that provide indicators, data sources, assumptions, or assessment methods transferable to educational contexts. By distinguishing between these two evidence streams, the review clarifies what is currently known about the environmental implications of generative AI in education and how future educational research could assess these impacts more systematically.

The review addresses the following research questions:

- RQ1: What types of environmental impacts are associated with the use of generative AI tools in educational contexts in the published literature?
- RQ2: What indicators, data sources, assumptions, and methodological approaches are used to measure, estimate, or report these environmental impacts?

To answer these questions, the synthesis distinguishes between education-specific evidence and complementary technical evidence from the broader AI sustainability literature.

2. Analytical Framework

This analytical framework treats the environmental footprint of generative artificial intelligence as a set of lifecycle and system-boundary questions rather than as a repeated inventory of electricity, water, hardware, and waste. Its purpose is to define how impacts are coded in the review: operational impacts, embodied and lifecycle impacts, end-of-life impacts, and levels of assessment. This distinction follows assessment work that separates computing-related emissions from wider system effects and stresses transparent reporting of the functional unit, hardware, location, energy source, and uncertainty [22–24].

2.1. Operational Environmental Impacts

Operational impacts occur during model training, fine-tuning, and inference. In this review, operational indicators are coded only when they are linked to identifiable AI workloads, such as training runs, deployed inference, prompt processing, image generation, or platform-level use, rather than to generic ICT energy use. The main indicators are electricity consumption, computational demand, greenhouse gas emissions, and water use associated with data center operation [4,12,16,25]. These impacts vary by model architecture, parameter count, hardware type, batch size, input-output ratio, task complexity, server utilization, data center efficiency, and the electricity mix used at the time and place of computation [13,26,27].

Training and inference are kept analytically separate because they produce different patterns of demand. Training concentrates computation during model development, whereas inference generates recurring demand whenever deployed systems are used [4,16,17]. This distinction is central for education, where most activity involves hosted inference through repeated prompts, revisions, translations, coding exchanges, image requests, feedback tools, and administrative applications [17–19]. A single interaction may be small, but course-, campus-, or system-level adoption can aggregate into material demand, especially when outputs require repeated verification or revision.

Operational carbon is therefore assessed as a function of workload, hardware, facility overhead, and grid carbon intensity, not as a fixed property of a prompt or model. Current energy-sector reporting emphasizes that AI-related data center demand must be interpreted alongside regional electricity supply, grid constraints, and the pace of infrastructure expansion [23]. For this reason, estimates that use tokens, prompts, or model size as proxies are treated as partial indicators unless they also specify the hardware, location, power usage effectiveness, and electricity-source assumptions used in the calculation.

Water use is coded separately from carbon because it follows different mechanisms and local constraints. Data centers may consume water directly through cooling systems and indirectly through electricity generation [4,15,16]. Mytton [28] shows that water reporting remains less mature than energy reporting and that direct consumption, potable-water sourcing, indirect electricity-related water use, and local water stress are not interchangeable measures. The framework therefore distinguishes direct facility water use, indirect water use associated with electricity generation, and cases where a study mentions water qualitatively without measuring it.

2.2. Embodied and Lifecycle Impacts

Environmental assessment limited to model operation captures only part of the total footprint. Embodied impacts result from raw material extraction, semiconductor production, manufacturing of graphics processing units and servers, construction of data center infrastructure, and transportation of equipment [13,15,16]. Lifecycle assessment broadens the system boundary by examining impacts across material extraction, production, transport, operation, maintenance, replacement, and disposal [13,21]. This approach is particularly relevant to generative AI because assessments based only on training or inference electricity may understate hardware manufacturing, facility construction, and supply-chain burdens.

Hardware and end-of-life impacts are coded as distinct lifecycle categories rather than folded into operational carbon. Wang et al. [29] estimate that generative AI could create a substantial additional e-waste stream by 2030 if server turnover accelerates, while the Global E-waste Monitor 2024 documents a wider gap between global e-waste generation and documented recycling [30]. These sources support treating processor replacement, server retirement, critical materials, hazardous components, and recycling limits as part of the assessment boundary. Educational institutions usually do not control these upstream and downstream processes directly, but their platform procurement, cloud contracts, and refresh cycles contribute to the demand signals that shape them.

2.3. Levels of Environmental Assessment

The environmental impacts of generative AI can be assessed at several analytical levels. At the hardware level, researchers measure processor power draw, utilization, memory use, and performance under controlled workloads [11,12]. At the model level, studies compare training energy, inference efficiency, parameter count, architecture, and emissions across systems [12,16]. At the task level, impacts may be expressed per prompt, token, generated image, page of text, coding solution, or completed activity [21,31–33]. At the user, classroom, or course level, researchers can combine activity logs with model-specific energy or emission estimates [18,19]. At the institutional level, assessment may include cumulative use across students, teachers, administrative services, learning platforms, and cloud contracts, together with procurement, server location, renewable-energy claims, platform architecture, and organizational carbon accounting [4,15,34].

These levels should not be treated as interchangeable. A prompt-level estimate cannot represent an institution-wide service without assumptions about frequency, user behavior, model routing, output quality, and retry rates; conversely, a hardware benchmark may not reflect actual educational usage patterns. Consistent with reporting guidance for machine-learning energy and carbon accounting [22], the review records the functional unit, system boundary, model, hardware assumptions, workload, electricity source, data center efficiency, and uncertainty wherever these are available.

2.4. Application of the Framework to the Review

This framework guides the review in two ways. First, it supports classification of the environmental impacts reported in the included studies without repeating the introductory background on electricity, water, hardware, and waste. Second, it provides a basis for comparing indicators, data sources, assumptions, system boundaries, and estimation methods across two evidence streams. Education-specific studies show how generative AI is used in teaching, learning, assessment, student work, and institutional settings [18–20]. Technical and environmental studies provide benchmarks, lifecycle methods, energy models, carbon accounting procedures, and software tools that may be applied to educational usage data [11–13,21]. Keeping these evidence streams analytically separate makes it possible to identify what has been measured directly in education and what remains dependent on estimates transferred from adjacent fields.

3. Materials and Methods

3.1. Research Design

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [35]. The PRISMA framework ensures

https://doi.org/10.3390/su18147213

a structured, transparent, and replicable process for identifying, screening, and selecting relevant studies, thereby enhancing the methodological rigor and credibility of the review findings. The review process included three key stages: (1) a comprehensive literature search across four academic databases (Scopus, Web of Science Core Collection, ERIC, and IEEE Xplore) to capture the interdisciplinary scope of the topic spanning education, generative artificial intelligence, and environmental sustainability; (2) a structured two-stage screening procedure based on predefined inclusion and exclusion criteria; and (3) a detailed full-text analysis and thematic synthesis of the included studies. Because the review addresses an emerging topic at the intersection of educational technology and environmental sustainability, the design incorporated two complementary evidence streams: education-specific publications examining the environmental impacts of generative AI, and contributions from the broader AI sustainability literature providing methodological approaches, indicators, and data sources applicable to educational contexts. This dual-purpose design is reflected in the search strategy and eligibility criteria and guided the thematic organization of the synthesis.

An internal review protocol was developed before the search and screening process, specifying the research questions, eligibility criteria, search strategy, data extraction form, quality appraisal framework, and synthesis plan. The protocol was not formally registered on a public registry because this review focuses on environmental rather than health outcomes. The extraction sheet and full-text exclusion list are provided as Supplementary files to support reproducibility. The PRISMA 2020 Checklist [35] for this review is provided as Supplementary Material (see: Table S4. PRISMA_2020_checklist).

3.2. Eligibility Criteria

Eligibility criteria were established before screening and aligned with the review aim and research questions. Because the review examines both evidence generated within educational contexts and measurement approaches developed in related technical fields, eligible publications were classified into two evidence streams: education-specific studies and complementary technical studies. This distinction was maintained during study selection, data extraction, and synthesis.

3.2.1. Education-Specific Studies

Studies were included in the education-specific evidence stream when they met all of the following criteria:

- Generative AI focus. The publication examined generative artificial intelligence, including large language models, AI chatbots, text-to-image systems, code-generating models, or other systems capable of generating new content.
- Educational context. The study addressed teaching, learning, assessment, curriculum design, student support, teacher work, professional development, educational administration, institutional implementation, or educational policy.
- Environmental relevance. The publication measured, estimated, reported, or substantively discussed at least one environmental impact associated with generative AI. Eligible impacts included electricity consumption, computational demand, greenhouse gas emissions, carbon footprint, data-center use, water use, hardware or infrastructure requirements, raw-material use, embodied impacts, and electronic waste.
- Relevance to the review questions. The study provided evidence concerning either the type of environmental impact associated with educational GenAI use or the indicators, data sources, assumptions, system boundaries, or methodological approaches used to assess that impact.

3.2.2. Complementary Technical Studies

Studies outside a direct educational context were included in the complementary technical evidence stream when they met all of the following criteria:

- Generative AI focus. The publication examined identifiable generative AI models, generative AI workloads, or infrastructure used specifically for generative AI.
- Transferable environmental evidence. The study reported an environmental indicator, measurement, model, dataset, calculation procedure, or assessment method that could be applied to educational GenAI use.
- Defined unit or method of assessment. The publication provided a transferable unit of analysis or methodological procedure, such as energy per query, token, image, task, model, or training run; emissions per functional unit; hardware-level power measurements; water-use estimates; lifecycle assessment; carbon-accounting methods; or software tools for estimating environmental impact.
- Relevance to educational assessment. The reported evidence could reasonably support estimation or interpretation of environmental impacts arising from student, teacher, classroom, course, platform, or institutional use of GenAI.

Technical studies were not included solely because they discussed the sustainability of digital technology or artificial intelligence in general. They had to provide GenAI-specific evidence or methods with a clear potential application to educational use.

3.2.3. Common Publication Criteria

For both evidence streams, publications were eligible when they:

- Were published between January 2022 and June 2026;
- Were written in English;
- Were available as full text;
- Were published as peer-reviewed journal articles, conference papers, or review articles.

The publication period was selected because the widespread public adoption of large language model-based generative AI tools began in late 2022, followed by rapid expansion of their use in education. Although some foundational lifecycle-assessment and carbon-accounting methods predate this period, the review focuses on GenAI-specific evidence; foundational methods that are not specific to generative AI were excluded unless they were cited or applied within an included GenAI-focused publication. This decision is a limitation, as some transferable methodological frameworks (e.g., general LCA standards) developed before 2022 may be relevant but are not reviewed directly.

3.2.4. Exclusion Criteria

Publications were excluded when they met any of the following conditions:

- They addressed artificial intelligence, machine learning, learning analytics, adaptive systems, or educational technology without a clear focus on generative AI.
- They discussed generative AI but contained no relevant information about environmental impacts, environmental indicators, computational demand, infrastructure, resource use, or assessment methods.
- They referred to sustainability only in a social, economic, ethical, pedagogical, or institutional sense without addressing an environmental dimension.
- They were outside education and did not provide transferable environmental evidence, indicators, data sources, assumptions, or methods applicable to educational GenAI use.
- They addressed general data-center sustainability or conventional computing without a clear connection to generative AI.

https://doi.org/10.3390/su18147213

- They were editorials, opinion pieces, news items, blog posts, book reviews, abstracts, posters, presentation slides, or other non-peer-reviewed publication types.
- The full text was unavailable.
- They were not published in English or fell outside the specified publication period.

When eligibility remained uncertain, the full text was assessed to determine whether the study contributed evidence relevant to at least one research question. Studies were assigned to only one evidence stream, based on whether their primary context was educational or technical.

3.3. Information Sources

The literature search was conducted in Scopus, Web of Science Core Collection, ERIC, and IEEE Xplore. These databases were chosen to ensure comprehensive coverage of the interdisciplinary review topic, which includes education, generative artificial intelligence, and environmental sustainability. Scopus and Web of Science offer broad multidisciplinary coverage of peer-reviewed research, ERIC focuses on educational literature, and IEEE Xplore indexes technical and engineering research related to artificial intelligence, computational infrastructure, and energy-efficient technologies. In addition to database searches, the reference lists of included review articles were examined to identify relevant primary or technical publications not retrieved through the database search strategy. This supplementary search was especially important for identifying GenAI-specific environmental measurement studies within the broader AI sustainability literature. Through this reference-list search, 18 additional records were identified. After deduplication against the database records, 12 remained for screening. Of these, 7 were excluded at title/abstract screening and 5 proceeded to full-text assessment. Three of these were included in the review, bringing the total from 20 database-derived included studies to 23. The supplementary-search pathway is reported separately in the revised PRISMA flow diagram (Figure 1).

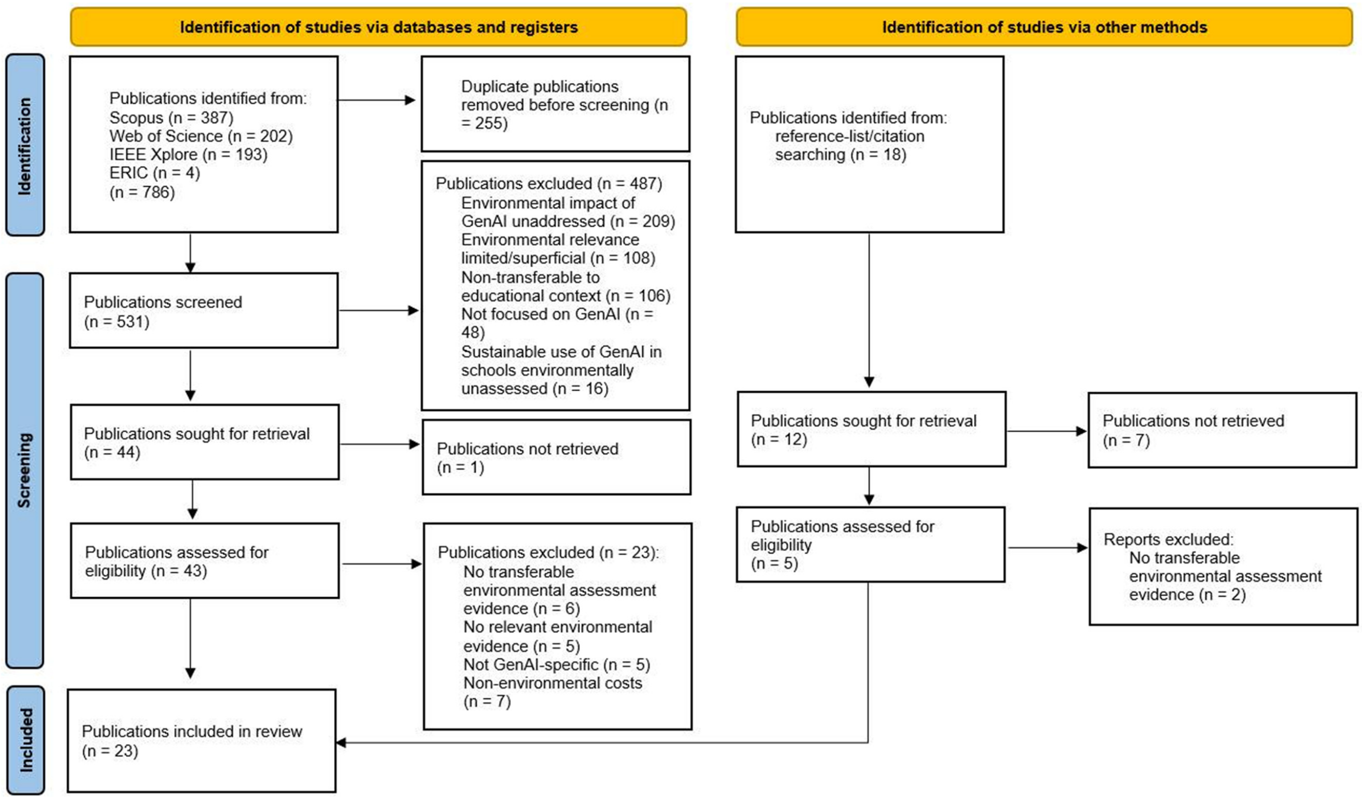

Figure 1. PRISMA flow diagram of systematic review.

3.4. Search Strategy

A systematic literature search was conducted to identify studies examining the environmental impacts of generative artificial intelligence (GenAI) in educational contexts. The search strategy was developed iteratively through preliminary scoping searches and refined to balance sensitivity and specificity. The search focused on three core concepts: (1) generative artificial intelligence, (2) education, and (3) environmental sustainability and environmental impacts. Keywords related to generative AI included terms such as generative AI, generative artificial intelligence, ChatGPT, large language model, LLM, AI chatbot, foundation models, GPT, Copilot, Gemini, diffusion models, text-to-image, image generation, and inference. Educational terms included education, teaching, learning, student, teacher, and educator. Environmental terms included environmental impact, carbon footprint, carbon emissions, energy consumption, computational cost, green AI, sustainable AI, water footprint, water consumption, life cycle assessment, hidden emissions, electronic waste, hardware, GPU, data center, cooling, supply chain, electricity use, and greenhouse gas.

Because the topic is relatively new and interdisciplinary, the search strategy prioritized broad retrieval of studies addressing generative AI in educational settings. Pilot searches indicated that a highly restrictive combination of educational and environmental search terms could exclude potentially relevant studies that discussed environmental implications only in the full text or discussion sections. The final search strategy combined three concept blocks covering generative AI, education, and environmental impacts. The terminology within each block was kept sufficiently broad to retrieve publications that discussed environmental implications at different levels of specificity. Detailed relevance to the review questions was subsequently assessed during title and abstract screening and full-text assessment. Search strings were adapted to the indexing structure and search functionalities of each database. The search was limited to publications published between January 2022 and June 2026. This period was selected because the widespread adoption of generative AI in education emerged after the public release and rapid diffusion of large language model-based tools in late 2022. Only publications written in English were considered. Peer-reviewed journal articles, conference papers, and review articles were eligible for inclusion. The complete search strings for each database are reported in Table 1. The searches were conducted on 6 June 2026. Search results from all databases were exported in RIS format and imported into Rayyan for deduplication and screening. Rayyan is a collaborative tool designed specifically for systematic literature reviews [36].

Table 1. Database Search Strategies and Records Retrieved.

Database	Search Query	Records Retrieved
Scopus	TITLE-ABS-KEY ((“generative AI” OR “generative artificial intelligence” OR ChatGPT OR “large language model” OR LLM OR “AI chatbot” OR “foundation model” OR GPT OR Copilot OR Gemini OR “diffusion model” OR “text-to-image” OR “image generation” OR inference) AND (“higher education” OR “educational context” OR “educational setting” OR “educational technolog” OR “teaching” OR “teacher” OR “educator” OR “student” OR “classroom” OR “curriculum” OR “assessment” OR “teacher education” OR “medical education”) AND (“carbon footprint” OR “carbon emission” OR “energy consumption” OR “computational cost” OR “environmental impact” OR “environmental cost” OR “water footprint” OR “water consumption” OR “life cycle assessment” OR “electronic waste” OR “greenhouse gas” OR “e-waste”)) AND PUBYEAR > 2021 AND PUBYEAR < 2027 AND LANGUAGE(English) AND (LIMIT-TO (DOCTYPE, “ar”) OR LIMIT-TO (DOCTYPE, “cp”) OR LIMIT-TO (DOCTYPE, “re”))	n = 387

https://doi.org/10.3390/su18147213

Table 1. *Cont.*

Database	Search Query	Records Retrieved
Web of Science Core Collection	TS = ((“generative AI” OR “generative artificial intelligence” OR ChatGPT OR “large language model” OR LLM OR “AI chatbot” OR “foundation model” OR GPT OR Copilot OR Gemini OR “diffusion model” OR “text-to-image” OR “image generation” OR inference) AND (“higher education” OR “educational context” OR “educational setting” OR “educational technolog” OR teaching OR teacher* OR educator* OR student* OR classroom* OR curriculum OR assessment OR “teacher education” OR “medical education”) AND (“carbon footprint” OR “carbon emission” OR “energy consumption” OR “computational cost” OR “environmental impact” OR “environmental cost” OR “water footprint” OR “water consumption” OR “life cycle assessment” OR “electronic waste” OR “greenhouse gas” OR “e-waste”))	*n* = 202
ERIC	(“generative AI” OR “generative artificial intelligence” OR ChatGPT OR “large language model” OR “large language models” OR LLM OR LLMs OR “AI chatbot” OR “AI chatbots” OR “foundation model” OR “foundation models” OR GPT OR Copilot OR Gemini OR “diffusion model” OR “diffusion models” OR “text-to-image” OR “image generation” OR inference) AND (“higher education” OR “educational context” OR “educational contexts” OR “educational setting” OR “educational settings” OR “educational technology” OR teaching OR teacher OR teachers OR educator OR educators OR student OR students OR classroom OR classrooms OR curriculum OR assessment OR “teacher education” OR “medical education”) AND (“carbon footprint” OR “carbon emission” OR “carbon emissions” OR “energy consumption” OR “computational cost” OR “computational costs” OR “environmental impact” OR “environmental impacts” OR “environmental cost” OR “environmental costs” OR “water footprint” OR “water consumption” OR “life cycle assessment” OR “electronic waste” OR “greenhouse gas” OR “e-waste”)	*n* = 4
IEEE Xplore	((“All Metadata”:“generative AI” OR “All Metadata”:“generative artificial intelligence” OR “All Metadata”:“ChatGPT” OR “All Metadata”:“large language model” OR “All Metadata”:“large language models” OR “All Metadata”:“LLM” OR “All Metadata”:“LLMs” OR “All Metadata”:“AI chatbot” OR “All Metadata”:“AI chatbots” OR “All Metadata”:“foundation model” OR “All Metadata”:“foundation models” OR “All Metadata”:“GPT” OR “All Metadata”:“Copilot” OR “All Metadata”:“Gemini” OR “All Metadata”:“diffusion model” OR “All Metadata”:“text-to-image” OR “All Metadata”:“image generation” OR “All Metadata”:“inference”) AND (“All Metadata”:“higher education” OR “All Metadata”:“educational context” OR “All Metadata”:“educational setting” OR “All Metadata”:“educational technology” OR “All Metadata”:“teaching” OR “All Metadata”:“teacher” OR “All Metadata”:“teachers” OR “All Metadata”:“educator” OR “All Metadata”:“educators” OR “All Metadata”:“student” OR “All Metadata”:“students” OR “All Metadata”:“classroom” OR “All Metadata”:“curriculum” OR “All Metadata”:“assessment” OR “All Metadata”:“teacher education” OR “All Metadata”:“medical education”) AND (“All Metadata”:“carbon footprint” OR “All Metadata”:“carbon emission” OR “All Metadata”:“carbon emissions” OR “All Metadata”:“energy consumption” OR “All Metadata”:“computational cost” OR “All Metadata”:“computational costs” OR “All Metadata”:“environmental impact” OR “All Metadata”:“environmental impacts” OR “All Metadata”:“environmental cost” OR “All Metadata”:“environmental costs” OR “All Metadata”:“water footprint” OR “All Metadata”:“water consumption” OR “All Metadata”:“life cycle assessment” OR “All Metadata”:“electronic waste” OR “All Metadata”:“greenhouse gas” OR “All Metadata”:“e-waste”))	*n* = 193

It should be noted that the main database search strategy required all three concept blocks (GenAI, education, and environment) to appear simultaneously, which was appropriate for identifying education-specific studies. Complementary technical studies that did not include educational terms were identified primarily through the supplementary reference-list search of included review articles. This dual-pathway approach—a structured database search for education-specific evidence and a reference-list search for transferable technical evidence—was chosen because a fully separate technical-database search (combining GenAI and environmental terms without the education block) would have returned an unmanageably large and largely irrelevant set of records for the specific purpose of this review. The supplementary search pathway is documented transparently in the PRISMA flow diagram and in the methods section.

3.5. Study Selection

Rayyan's duplicate-detection function initially identified 198 duplicate records. An additional 57 records with matching titles and authors were identified during manual checking and confirmed as duplicates. In total, 255 duplicate records were removed before eligibility screening. No records were excluded solely through an automated eligibility decision.

Study selection proceeded in two stages. In the first stage, one reviewer screened the titles and abstracts of all remaining records against the predefined eligibility criteria. Records were retained if they appeared relevant to either of the two evidence streams. Education-specific records were expected to address generative AI, an educational context, and an environmental dimension. Records without a direct educational context were retained if they appeared to provide GenAI-specific environmental indicators, measurements, data sources, assumptions, benchmarks, or assessment methods that could be transferred to educational use. The 106 records excluded at the title/abstract stage for "non-educational context" were those whose titles and abstracts indicated no connection to generative AI (e.g., general machine learning, conventional computing, or non-AI educational technology) and no transferable environmental evidence. The 15 complementary technical studies included at full-text stage were identified because their abstracts or full texts explicitly addressed GenAI-specific environmental measurement, benchmarking, or lifecycle assessment with clear transferability to educational use. The operational distinction is thus: exclusion for "non-educational context" applied when a record lacked both educational relevance and GenAI-specific transferable evidence; inclusion as a complementary technical study required GenAI-specific environmental evidence or methods even without an educational setting. Records were excluded at the title and abstract stage if they clearly lacked a generative AI focus, contained no environmental dimension, addressed sustainability only in a social, ethical, economic, institutional, or pedagogical sense, fell outside the eligible publication period or publication types, or were unrelated to either evidence stream. When the title and abstract did not provide enough information to determine eligibility, the record was retained for full-text assessment.

In the second stage, the full texts of potentially eligible publications were assessed against the complete eligibility criteria. Education-specific studies were included if they identified, measured, estimated, reported, or substantively discussed an environmental impact associated with GenAI use in education. Complementary technical studies were included if they provided an environmental indicator, benchmark, data source, assumption, calculation procedure, lifecycle approach, or other assessment method that could reasonably inform the evaluation of educational GenAI use. Review articles were retained if they provided substantive secondary evidence relevant to at least one research question and met the publication-type criteria.

Screening was conducted by two reviewers. In the first stage (title and abstract screening), one reviewer initially screened all records, while the second reviewer independently double-screened a random 20% sample (106 records) to check consistency. The two reviewers agreed on 98 of these 106 records (agreement rate: 92.5%). The eight disagreements were resolved through discussion until consensus was reached; no third reviewer was required. In the second stage (full-text assessment), both reviewers independently assessed all 43 full-text articles. Initial agreement was reached on 39 of 43 articles (agreement rate: 90.7%). The four disagreements were resolved through discussion, and a single primary reason for exclusion was recorded for each excluded publication. No formal inter-rater reliability statistic (e.g., Cohen's kappa) was calculated, but agreement rates are reported here for transparency. A pilot calibration exercise was conducted on 20 records before the main screening to align the application of eligibility criteria. Data extraction, coding, and quality appraisal were conducted by one reviewer due to resource constraints; however, all uncertain or borderline classifications were revisited and checked by the second reviewer. This limitation is acknowledged in the discussion.

A single primary reason was recorded for each publication excluded after full-text assessment. Full-text exclusion reasons included lack of a clear GenAI focus, absence of relevant environmental evidence, no educational context or transferable technical relevance, a non-environmental conception of sustainability, ineligible publication type, publication outside the specified period, publication in a language other than English, or unavailable full text. Publications were not excluded solely due to methodological limitations identified during quality appraisal. Publications identified through the reference lists of included review articles were assessed separately. Each publication was first checked for duplication against the database records and then evaluated using the same eligibility criteria applied during full-text screening. Publications without a direct educational context were included only if they provided GenAI-specific environmental indicators, measurements, data sources, assumptions, benchmarks, or assessment methods that could reasonably inform the assessment of educational GenAI use. The search and selection process is presented in the PRISMA 2020 flow diagram (Figure 1). Following full-text assessment, 23 publications met the eligibility criteria and were included in the qualitative synthesis.

3.6. Data Extraction

Data were extracted from all publications that met the eligibility criteria after full-text screening. A structured extraction form consisting of five questions was developed in Rayyan and aligned with the review aim and research questions. The same extraction questions were applied to education-specific publications, complementary technical publications, and review articles.

1. The first question recorded the publication type. Publications were classified as journal articles, conference papers, review articles, or another eligible publication category.
2. The second question recorded the educational context in which generative AI was discussed or used. Relevant contexts included higher education, primary or secondary education, teacher education, professional development, medical education, institutional or policy settings, and other educational contexts. For complementary technical publications without a direct educational setting, the item was marked as not applicable, and the potential relevance of the publication to educational assessment was described in the accompanying notes.
3. The third question identified the type of GenAI use examined in the publication. Categories included teaching, learning, assessment, student support, teacher work, curriculum development, content generation, educational administration, and institutional implementation. More than one category could be selected. For technical publications, the notes described how the model, workload, benchmark, or measurement method could relate to student, teacher, course, platform, or institutional use of GenAI.
4. The fourth question captured the types of environmental impact identified or discussed. Categories included energy or electricity consumption, greenhouse gas emis-

sions, carbon footprint, computational demand, data-center operation, water consumption, hardware and infrastructure requirements, raw-material use, embodied or lifecycle impacts, and electronic waste. Multiple environmental impact categories could be assigned to a single publication.

5. The fifth question recorded the measurement or estimation approach used to assess or report environmental impact. Relevant approaches included direct hardware measurement, platform or application logs, model benchmarking, token-based estimation, secondary-data estimation, carbon-accounting calculations, lifecycle assessment, software-based estimation tools, survey-based evidence, and conceptual discussion without original quantification. More than one approach could be recorded where appropriate.

Additional information needed for the synthesis was recorded in accompanying extraction notes. These notes included the disciplinary field and study aim, the GenAI model or system examined, environmental indicators and functional units, data sources, hardware and infrastructure characteristics, system boundaries, electricity grid carbon intensity, power usage effectiveness, conversion factors, key assumptions, main findings relevant to the review questions, reported limitations, and the potential transferability of technical evidence to educational contexts. Information that was missing or unclear was recorded as “not reported” rather than inferred, while items that did not apply to a particular publication were marked as “not applicable.” One reviewer completed the extraction directly in Rayyan. Uncertain or internally inconsistent entries were revisited during the analysis and checked against the full text. Study authors were not contacted to obtain additional information.

3.7. Data Synthesis

The analysis proceeded in three stages.

Stage 1: Classification of study contexts and environmental impacts.

The included publications were first classified by evidence stream, publication category, disciplinary field, study aim, educational context, and type of GenAI use. Environmental impacts were then coded into categories including energy and electricity consumption, greenhouse gas emissions, carbon footprint, computational demand, data-center operation, water consumption, hardware and infrastructure requirements, raw-material use, embodied lifecycle impacts, and electronic waste. Publications could be assigned to more than one environmental impact category. For education-specific studies, the impacts were also linked to the educational activity examined, such as student prompting, teacher use, assessment, curriculum development, learning-platform integration, or institutional deployment.

Stage 2: Comparison of indicators and methodological approaches.

The second stage examined how environmental impacts were measured, estimated, or discussed. Publications were compared based on the indicators and functional units used, data sources cited, assumptions reported, system boundaries defined, and calculation or estimation procedures applied. Methodological approaches were grouped into categories such as direct hardware measurement, model benchmarking, platform or application logs, token-based estimation, secondary data estimation, carbon accounting models, lifecycle assessment, software-based estimation tools, survey-based evidence, and conceptual discussion without original quantification. The degree of empirical grounding was also considered. Evidence was classified into five categories: (1) directly measured—original hardware-level or platform-level measurements collected by the study authors; (2) estimated from primary activity data—original usage data (e.g., token counts, prompt logs) combined with secondary emission factors; (3) modelled from secondary assumptions—lifecycle or emissions estimates based entirely on published coefficients and assumptions without original usage data; (4) review-based—findings drawn from secondary synthesis of prior studies; and (5) conceptual commentary—discussion without original quantification. Review articles were treated as secondary evidence used to contextualize findings rather than as independent primary evidence. This distinction was used to interpret the specificity and credibility of the findings rather than to assign formal numerical weights to individual studies.

Stage 3: Comparison of evidence streams and identification of research gaps.

The final stage compared education-specific studies with complementary technical studies. The comparison focused on the types of environmental impacts examined, the depth of measurement, the indicators and methods used, and the extent to which technical approaches could be transferred to educational settings. Review articles were treated as secondary evidence and used to contextualize broader patterns. Quantitative estimates reported in review articles were not counted as independent primary findings when the corresponding original study was also included. The synthesis was presented narratively and supported by comparative tables. These tables summarized study characteristics, educational contexts, environmental impact categories, indicators, measurement approaches, key findings, and reported limitations. Frequencies and percentages were calculated for descriptive purposes where categories could be applied consistently. Because individual publications could address more than one environmental impact or methodological approach, category totals were not always mutually exclusive. Research gaps were identified by examining which environmental impacts and educational contexts were frequently or rarely addressed and which methodological details were absent or insufficiently reported. Particular attention was given to missing information concerning model identity, hardware, functional units, system boundaries, electricity-grid carbon intensity, data-center efficiency, water use, embodied impacts, and uncertainty. One reviewer conducted the coding and synthesis. Uncertain classifications were revisited and checked against the full texts and predefined category definitions.

3.8. Quality Assessment

The methodological quality and reporting transparency of the included publications were assessed to support interpretation of the synthesis. Because the review included empirical educational studies, technical and modeling studies, and review articles, no single appraisal tool was suitable for all publication types. Therefore, publication-type-specific appraisal criteria were applied. Empirical qualitative, quantitative, and mixed-methods studies were assessed using the Mixed Methods Appraisal Tool (MMAT; [37]). The MMAT provides separate sets of criteria for qualitative research, randomized controlled trials, non-randomized studies, quantitative descriptive studies, and mixed-methods research. Each empirical study was assessed against the five criteria corresponding to its methodological category. Criteria were rated as "yes", "no", or "cannot tell." Overall numerical scores were not calculated, as the MMAT guidance discourages combining individual criteria into a single summary score. Technical and modeling studies that did not correspond to an MMAT category were assessed using a review-specific framework developed for this study. The framework examined: (1) clarity of the study aim, environmental outcome, and unit of analysis; (2) reporting of the GenAI model, hardware, workload, and system boundaries; (3) appropriateness and traceability of the data sources; (4) transparency of assumptions, conversion factors, allocation rules, and calculation procedures; (5) consideration of uncertainty, sensitivity, limitations, or potential sources of error; and (6) reproducibility or transferability of the method. Each criterion was rated as "yes", "partly", "no", or "not applicable."

Review articles were assessed separately according to: (1) clarity of the review objective or question; (2) transparency of the information sources and search procedures; (3) specification of inclusion and exclusion criteria; (4) description of the study-selection or evidence-selection process; (5) appropriateness and clarity of the synthesis method; and (6) traceability of the conclusions to the cited evidence and acknowledgment of limitations. Narrative reviews that did not report reproducible search and selection procedures were retained as contextual secondary evidence when they met the eligibility criteria, but their findings were interpreted more cautiously than those of systematic reviews or primary empirical studies. Quality appraisal did not determine study inclusion. No publication was excluded solely because of its appraisal result. Instead, the assessment distinguished between evidence based on direct measurement, transparent modeling, secondary estimation, and conceptual discussion. Interpretation placed greater weight on studies that clearly reported their system boundaries, data sources, assumptions, functional units, and uncertainty. Claims based on incomplete reporting, heterogeneous secondary estimates, or non-systematic synthesis were treated with greater caution. To ensure transparency and repeatability, the topic-level "strength of evidence" labels used in the synthesis were defined as follows: "Strong" evidence was assigned when at least two independent studies using direct measurement or controlled experimentation reported consistent findings with transparent reporting of model, hardware, functional unit, system boundary, and uncertainty. "Moderate" evidence was assigned when findings were supported by either a single high-quality direct measurement or multiple studies using estimation or modeling with partially transparent assumptions, or when findings were consistent across studies but based on indirect proxies. "Limited" evidence was assigned when findings relied on secondary estimates, conceptual discussion, or a single study without independent replication, or when key methodological details (e.g., system boundary, functional unit, hardware assumptions) were missing or unclear.

One reviewer conducted the quality assessment for all 23 included publications. To improve consistency, all ratings were revisited after the initial appraisal and checked against the full texts, extraction records, and predefined appraisal criteria. We acknowledge that single-reviewer data extraction, coding, synthesis, and quality appraisal introduces a risk of subjective bias, particularly for a review involving heterogeneous evidence streams and multiple environmental indicators. The second reviewer independently verified extraction and appraisal for all uncertain or borderline cases, but a full double-extraction and double-appraisal process was not feasible within the available resources. This limitation is discussed in the Conclusions. When available information was insufficient to support a definitive judgment, ratings were recorded as "cannot tell" or "partly" rather than inferred. The complete appraisal criteria are presented in Supplementary Table S1, and the study-level appraisal results are reported in Supplementary Table S2.

3.9. Use of Generative AI

During the preparation of this manuscript, the authors used ChatGPT (OpenAI; GPT-5.5) for the purposes of GenAI-assisted translation, editing of the text and image creation (Figure 2, the policy framework diagram). The authors have reviewed and ed-ited the output and take full responsibility for the content of this publication.

4. Results

This systematic review analyzed and synthesized evidence on the environmental impacts of generative artificial intelligence (GenAI) in educational contexts, using both education-specific research and related technical and comparative studies. A total of 23 publications, comprising empirical studies, technical benchmarks, and review articles, representing the state of the art from 2022 to 2026 were fully analyzed. To address the core review questions (RQ 1 and RQ 2), the findings are organized into six key themes: (a) operational electricity consumption and grid demand, (b) operational carbon and greenhouse-gas footprints, (c) water consumption and data-center cooling demands, (d) embodied resources, supply chains, and e-waste lifecycle impacts, (e) comparative labor footprints—human versus artificial intelligence, and (f) educational interventions, student awareness, and behavioral feedback. These themes were generated through an iterative thematic analysis process: after initial coding of environmental impact categories (Stage 1 of synthesis), the first author grouped coded impacts into preliminary clusters based on shared environmental indicators and methodological approaches. These clusters were then refined through comparison across studies and discussion among all authors, resulting in the six final themes. The themes were not predefined but emerged from the data; alternative groupings (e.g., by analytical level or evidence stream) were considered but rejected because the thematic organization best reflected the dominant patterns in the literature.

Table 2 summarizes the main characteristics of the 23 included publications, distinguishing between education-specific studies and complementary technical publications used to support the assessment of environmental impacts in educational GenAI contexts.

The included publications show a clear imbalance in the current evidence base. Only eight publications examined GenAI in direct educational settings, mainly higher education, medical education, engineering education, design education, and sustainability education, whereas fifteen publications came from complementary technical fields such as computer science, environmental engineering, green AI, lifecycle assessment, and technology policy. This distribution confirms that education-specific evidence remains limited and still depends heavily on methods, indicators, and assumptions developed outside educational research. Empirical educational studies mostly addressed student use, prompting behavior, awareness, or platform-level interventions, while technical publications provided the more detailed evidence on energy consumption, carbon accounting, water use, lifecycle impacts, hardware efficiency, and data-center assumptions. This distinction was therefore retained throughout the synthesis. In each thematic section, we explicitly label findings as "observed in educational settings" when they derive from direct measurement or estimation in educational contexts, or as "inferred from technical literature" when they are drawn from hardware measurements, lifecycle simulations, or benchmarking studies outside education and transferred to the educational context by analogy. This labeling is intended to prevent the conflation of technical-simulation results with direct educational observations.

A detailed study-level evidence matrix (Table S3) is provided in the Supplementary Materials. For each included study, the matrix reports: the model or system examined; the task type and educational activity; the functional unit; the environmental outcomes assessed; data sources; hardware and server information; power usage effectiveness (PUE); regional power carbon intensity; the water consumption coefficient; geographical and electricity-grid assumptions; the system and lifecycle boundary; whether training emissions are allocated; whether hardware (embodied) emissions are included; the method for handling or reporting uncertainty; the type of evidence (direct measurement, estimated from primary data, modelled from secondary assumptions, review-based, or conceptual); and an appraisal outcome noting key strengths and limitations. This evidence matrix supports the auditing of environmental claims and methodological assumptions reported in the synthesis.

https://doi.org/10.3390/su18147213

Table 2. Characteristics of Included Publications.

Study	Year	Evidence Stream	Publication Type/Design	Main Field	Country/Region	Educational Context or Technical Relevance	GenAI System/Use
Dungo et al. (2025) [38]	2025	Education-specific	Journal article; quantitative descriptive survey	Educational technology/environmental studies	Philippines	Higher education; college students	General student use of GenAI for productivity and entertainment
Ansari et al. (2026) [39]	2026	Education-specific	Journal article; mixed-methods survey and topic analysis	AI ethics/digital sociology/environmental studies	India	Higher education; technical university students	ChatGPT, Gemini, Perplexity AI, DeepSeek, and Llama used for general productivity
Andersen et al. (2026) [18]	2026	Education-specific	Conference paper; non-randomized experimental study	HCI/responsible AI/educational technology	Germany	Undergraduate statistics course	GPT-4o, GPT-4o-mini, and GPT-4.1 used for study support with real-time CO_2 feedback
Esho et al. (2026) [16]	2026	Complementary technical	Review article	Environmental sustainability/green computing/AI	International/not specified	No direct educational setting; transferable sustainability evidence	Generative AI and quantum computing reviewed comparatively
Marmouzi et al. (2026) [17]	2026	Complementary technical	Systematic review	Green AI/sustainable AI/computer science	International	No direct educational setting; taxonomy and metrics transferable to education	AI systems including LLMs and other generative models
Wen et al. (2026) [40]	2026	Education-specific	Research article; prototype validation	Computer science/educational technology	China	Digital learning platform infrastructure	GAIA framework uses lightweight AI/ML models as an energy-efficient alternative to large LLMs
Rizzo (2025) [34]	2025	Complementary technical	Journal article; conceptual and analytical study	Electrical engineering/sustainable energy	International/not specified	No direct educational setting; general environmental implications of AI	AI and LLMs discussed in relation to sustainable energy systems
Ren et al. (2024) [31]	2024	Complementary technical	Journal article; empirical comparative study	Computer science/environmental studies	International/comparative	No direct educational setting; knowledge-work comparison relevant to education	LLMs compared with human labor across environmental indicators
Caravaca (2026) [11]	2026	Complementary technical	Conference/workshop paper; empirical benchmarking	Computer science	International/technical benchmarking	No direct educational setting; inference benchmarks transferable to education	Multiple LLMs, including DeepSeek and Llama models, tested across GPUs
Wu et al. (2026) [15]	2026	Complementary technical	Journal article; conceptual framework and analytical study	Computer science/sustainability	International/not specified	No direct educational setting; infrastructure-level framework	AI, big data, and ICT analyzed through green/red AI concepts
Sivapragasam et al. (2026) [41]	2026	Education-specific	Journal article; quantitative descriptive survey	Engineering education	India	Higher education; undergraduate engineering students	ChatGPT and LLMs used for writing, coding, problem-solving, exam preparation, and brainstorming

Table 2. *Cont.*

Study	Year	Evidence Stream	Publication Type/Design	Main Field	Country/Region	Educational Context or Technical Relevance	GenAI System/Use
Jung et al. (2025) [19]	2025	Education-specific	Conference paper; experimental methodological study	Medical informatics/ medical education	South Korea/United States	Medical education	ChatGPT-based tasks; translation and prompt paraphrasing used to reduce token counts
Woo (2025) [33]	2025	Complementary technical	Journal article; comparative experimental/modeling study	Computer science/ computational sustainability	United States	Competitive programming; technically relevant to AI-assisted learning	GPT-based models compared with human programmers on correctness-controlled coding tasks
Mbah et al. (2025) [42]	2025	Education-specific	Review article; critical thematic review	Education/ sustainability studies	United Kingdom	Sustainability education	ChatGPT and related tools used for information search, explanation, summarizing, and writing support
Rincé et al. (2025) [13]	2025	Complementary technical	Software/method paper	Computer science/ environmental informatics	Belgium	No direct educational setting; tool applicable to educational GenAI use	EcoLogits estimates energy, carbon, water, and embodied impacts of GenAI API requests
Jetly et al. (2026) [43]	2026	Complementary technical	Conference paper; narrative review	Computer science/environmental sustainability	International/not specified	No direct educational setting; wider AI-sustainability evidence	ChatGPT and other LLMs reviewed in relation to energy, carbon, water, hardware, and e-waste
d'Orgeval et al. (2026) [21]	2026	Complementary technical	Journal article; lifecycle simulation study	Applied energy/ environmental engineering	France	No direct educational setting; infrastructure evidence transferable to education	GPT-4o, LLaMA 3.1 405B, and DeepSeek V3 across multiple data-center typologies
Luccioni et al. (2024) [12]	2024	Complementary technical	Conference paper; empirical benchmarking	Computer science/ AI accountability	International/ technical benchmarking	No direct educational setting; inference evidence transferable to education	Task-specific and general-purpose models compared over 1000 inferences
Bouza et al. (2023) [14]	2023	Complementary technical	Research article; methodological review and guide	Computer science/ environmental research	France/United Kingdom	No direct educational setting; carbon-estimation methods transferable to education	Deep-learning training tools and carbon-estimation methods
Tomlinson et al. (2023) [32]	2023	Complementary technical	Journal article; comparative emissions analysis	Computer science/ environmental sustainability	United States	No direct educational setting; creative-task comparison relevant to education	ChatGPT and DALL-E 2 compared with human writers and illustrators
Malik (2025) [44]	2025	Complementary technical	Lifecycle modeling study	Sustainable AI/ lifecycle assessment	Spain	No direct educational setting; LCA framework transferable to education	GPT-3, ChatGPT, GPT-4, LLaMA 2, PaLM 2, and DistilBERT
Lupetti et al. (2025) [20]	2025	Education-specific	Case study/critical reflection with energy estimation	Design education/ sustainable HCI	Netherlands	Higher education; design workshop with 49 students	Text- and image-generation tools used in a design workshop
Kneese et al. (2024) [7]	2024	Complementary technical	Critical essay/review	Technology policy/ environmental justice	United States	No direct educational setting; lifecycle and justice implications	Generative AI examined through carbon, infrastructure, and regulatory perspectives

4.1. Theme 1: Operational Electricity Consumption and Grid Demand

4.1.1. Synthesis

Running generative AI models requires a substantial amount of electricity, divided between model pre-training and operational inference [16,17,38]. During model training, technical studies report high fixed energy costs: training the GPT-3 model consumed 1287 MWh, while the BLOOM model required 50.5 tons CO_2e [16,39]. Once deployed for educational use, the cumulative energy consumed by thousands of daily student queries during the inference phase becomes the dominant operational load, with inference estimated to account for 80% to 90% of a model's lifetime energy consumption under specific assumptions about deployment scale and user volume reported in the cited technical studies [11,13]. This figure should not be generalized to all models or deployment scenarios without considering the specific conditions (model size, user population, interaction frequency) under which it was estimated. In academic settings, GenAI deployment significantly shifts local computing footprints. Lupetti et al. [20] tracked a design workshop where 49 students integrated text-to-image and text-to-text models into active learning assignments. They found that adopting generative AI tasks doubled the aggregate energy consumption of students' educational activities compared to traditional workstations without AI. To mitigate these central server demands, Wen and Lv [45] developed the GAIA optimization framework. By replacing cloud-hosted LLM pathways with lightweight, localized machine learning models optimized for campus servers, they reduced platform operational energy consumption by 30%, saving 80 kWh per run while maintaining responsive user speeds.

4.1.2. Methodological Comparison

Technical studies measure electricity consumption directly using hardware-level interfaces, such as the NVIDIA System Management Interface (NVML), to record actual GPU power draw in watts or joules during controlled testing runs [11,12]. Caravaca [11] analyzed over 30,000 empirical measurements across 50 GPU architectures, demonstrating that operational energy consumption scales exponentially with model parameter count. In contrast, educational studies cannot access the physical hardware of proprietary APIs, such as those hosted by OpenAI or Anthropic. Instead, they rely on indirect software calculators and API token logs [18,19]. While technical studies compare raw model structures on standardized benchmarks, educational research estimates campus energy consumption by translating student platform activity logs into estimated watt-hours. They do this by combining input and output token counts with public cloud efficiency reports or Power Usage Effectiveness (PUE) ratings, which typically range from 1.12 to 1.18 [13,18].

4.1.3. Strength of Evidence

The evidence linking GenAI use to increased electrical workloads is supported by direct hardware measurements, supported by high-precision physical measurements in engineering laboratories [11,12] and confirmed on local student networks [20]. Laboratory benchmarks consistently show that multi-purpose generative models are significantly more energy-intensive than task-specific, fine-tuned models performing the same tasks [11]. However, evidence regarding actual campus-wide energy changes is moderate. Most educational studies rely on general secondary statistics or assume a direct, linear relationship between token counts and energy consumption, overlooking system variables such as server utilization, concurrent query batches, and server idle-state power.

4.1.4. Research Gaps

Current literature does not measure real-time electricity consumption across entire university networks, relying instead on limited workshop data or isolated server prototypes. No studies examine how institutional academic calendars, exam periods, or campus-wide typing patterns influence localized utility grid spikes. Additionally, research has not addressed the energy differences between web-based chat interfaces and programmatic API integrations built directly into university learning management systems.

4.2. Theme 2: Operational Carbon and Greenhouse Gas Footprints

4.2.1. Synthesis

Inference-phase carbon emissions (measured in CO_2 or CO_2eq) depend on the operational energy consumed and the carbon intensity of the local electricity grid supporting the active data center [14,21]. While model training emits substantial greenhouse gases, such as 552 metric tons of CO_2eq for GPT-3 pre-training [16], daily student prompting generates emissions at the milligram scale per request [18]. Andersen et al. (2026) [18] tracked 781 student prompts in an active undergraduate statistics course and calculated model-specific carbon intensity. They found that average emissions per token were 0.198 mg CO_2eq for GPT-4o, 0.081 mg CO_2eq for GPT-4o-mini, and 0.299 mg CO_2eq for GPT-4.1. This shows that changing models can alter classroom emissions by up to 269% for equivalent academic tasks in the specific research environment. This figure reflects the specific models, tasks, and grid conditions studied and should not be generalized to other educational settings without replication. Additionally, full-architecture life cycle modeling confirms that server geography is the most critical determinant of carbon emissions [18]. Routing identical student queries to data centers powered by clean hydro or nuclear grids reduces net emissions by over 90% compared to fossil-heavy networks [13,14].

4.2.2. Methodological Comparison

Methodological approaches reveal a sharp divide between bottom-up lifecycle simulation models and token-based calculators. Carbon-accounting software packages such as EcoLogits [13] and Green Algorithms [14] incorporate active server specifications, regional grid carbon intensity, and real-time PUE estimates to calculate emissions. In contrast, some educational studies use prompt token counts as a direct proxy for carbon footprints without considering server conditions [19]. Although reducing token counts by translating queries from Korean to English and removing unnecessary words reduced total tokens by 20% [19], treating token counts as directly proportional to carbon emissions is methodologically incomplete. This approach ignores that different model architecture types (e.g., Mixture-of-Experts vs. dense models) activate different fractions of their total parameters per token, resulting in substantial variations in energy consumption.

4.2.3. Strength of Evidence

The evidence that model selection and regional grid location determine carbon footprints is supported by lifecycle modelling, supported by rigorous, peer-reviewed computational models and cloud-provider validation [13,14,19]. However, the educational literature's reliance on token counts as a direct proxy for carbon emissions remains a weak methodological choice. Most educational studies do not account for grid fluctuations, data center cooling overheads, or active query batching, reducing the accuracy of their reported student carbon estimates.

https://doi.org/10.3390/su18147213

4.2.4. Research Gaps

We found no empirical studies measuring the operational carbon footprint of generative AI tools in primary, secondary, or adult education contexts; the empirical literature is limited to higher education. There is also a complete lack of empirical research examining carbon emissions from high-volume administrative tasks, such as automated grading, admissions screening, or lecture transcription. Researchers have not evaluated whether student platforms can dynamically route API calls to low-carbon regions based on real-time grid conditions.

4.3. Theme 3: Water Consumption and Data-Center Cooling Demands

4.3.1. Synthesis

Operating the remote physical infrastructure required for GenAI consumes substantial amounts of water, both on-site through evaporative cooling towers (Scope 1) and off-site during electricity generation (Scope 2) [15,17,44]. We should note that the terms "Scope 1" and "Scope 2" are conventionally used for greenhouse-gas reporting rather than water accounting; water consumption is discussed here in terms of direct (on-site cooling) and indirect (electricity-generation) pathways without assigning formal scope labels. Global projections indicate that AI-related water demand will reach 4.2 to 6.6 billion cubic meters by 2027 [15,39]. At the user level, standard estimates cited in reviews suggest that a typical conversation of 10 to 50 prompts with ChatGPT was estimated to use approximately 500 mL of water, depending on data center PUE and local humidity, as reported in review studies drawing on corporate sustainability disclosures [38,43]. This estimate is modelled from secondary data rather than directly measured and reflects assumptions about specific data-center configurations; it should not be treated as a stable general estimate across all deployments or regions. These impacts are geographically unequal, with two-thirds of modern data centers built in water-stressed regions where cooling demand directly threatens local drinking water supplies [4,7].

4.3.2. Methodological Comparison

Technical and environmental reviews calculate water footprints using broad water-intensity coefficients (milliliters of water per kWh consumed) provided by cloud operators or derived from corporate sustainability report data [15,16]. Educational research has failed to produce primary empirical data for water metrics. Although surveys assess student concerns about water scarcity [38,41], no educational trials track student prompts or platform logs to estimate on-site cooling water use. Instead, educational studies copy generalized global water statistics from computer science reviews and apply them to teaching activities without adjusting for the specific hardware, server cooling technology, or regional water stress index of the active data center.

4.3.3. Strength of Evidence

The evidence linking the physical mechanisms of data center water consumption to the threat of regional water stress is moderate. Key calculations rely on generalized corporate disclosures and top-down municipal utility estimates [16,44]. In education-specific settings, the strength of evidence is limited. No primary empirical studies have directly validated water metrics in educational contexts, leaving the field reliant on coarse, non-localized estimates drawn from unrelated industrial settings.

4.3.4. Research Gaps

There is no research tracking localized water usage in educational technology portfolios. We found no studies mapping how the seasonal utility bills of universities or school districts change when digital learning resources shift from local files to generative cloud interfaces. Furthermore, research has not explored how to integrate localized water-stress indexes into educational AI platforms to warn students when their prompts are routed to water-scarce regions.

4.4. Theme 4: Embodied Resources, Supply Chains, and e-Waste Lifecycle Impacts

4.4.1. Synthesis

The environmental footprint of generative AI starts before model execution and extends beyond software retirement. Embodied impacts (Scope 3) include the carbon, water, and material footprints of raw mineral mining, semiconductor fabrication, processor assembly, and physical product transport [7,13,16]. Manufacturing high-performance graphics processing units (GPUs) requires rare-earth metals and ultra-pure water, resulting in high Abiotic Depletion Potential (ADP) and toxic chemical waste [16,17]. Whole-architecture life cycle assessments (LCAs) indicate that embodied manufacturing emissions are major contributors to AI's net footprint, adding a fixed 22% to the total operational training footprint of models such as BLOOM-176B [13]. Additionally, because GenAI servers operate at high thermal capacity under continuous multi-tenant workloads, high-end GPUs are estimated to experience accelerated obsolescence, with active lifespans estimated at 3–5 years under high-utilization multi-tenant workloads, though this range depends on workload intensity, cooling conditions, and replacement policies [16,43]. This rapid hardware replacement cycle produces substantial volumes of toxic electronic waste (e-waste), most of which is exported to vulnerable communities in the Global South [39,43].

4.4.2. Methodological Comparison

Technical environmental engineers use comprehensive Life Cycle Assessment (LCA) software frameworks linked to industrial material databases, such as Ecoinvent, to model Scope 3 impacts across the four manufacturing stages: raw material extraction, processing, assembly, and transportation [13,21]. In contrast, educational research omits embodied lifecycle stages entirely, focusing exclusively on operational electricity. Because educational institutions do not purchase physical GPU server clusters—relying instead on on-demand cloud software configurations—their internal procurement guidelines and green hardware codes do not account for the upstream mining or downstream e-waste generated by their software activities [20].

4.4.3. Strength of Evidence

The technical engineering evidence documenting high toxic waste and raw material demands during GPU development is strong, based on standard, globally recognized ISO 14040/44 [46,47] LCA validation metrics [13,21]. However, the educational literature addresses these factors only minimally. Most education journals discuss e-waste and supply chain justice conceptually, without translating upstream mining metrics into actionable campus procurement or digital use policies.

4.4.4. Research Gaps

No studies have investigated the total Scope 3 embodied carbon or e-waste footprint of a university's local IT operations before and after adopting cloud-based GenAI tools. Researchers have not evaluated whether local e-waste recycling rates change when school systems shift from hardware-intensive computing labs to browser-based, AI-driven learning tools. We also found no research exploring how to incorporate raw metal scarcity metrics or supplier environmental justice scores into educational AI purchasing frameworks.

4.5. Theme 5: Comparative Labor Footprints: Human vs. Artificial Intelligence

4.5.1. Synthesis

An emerging area of the literature compares the environmental footprint of human cognitive labor with automated machine computation [31–33]. Early calculations claimed that AI writing and illustrating generated far fewer emissions than human equivalents: ChatGPT was estimated to emit 130 to 1500 times less CO_2eq per page than a human writer, while DALL-E2 emitted 310 to 2900 times less than a human illustrator [32]. According to these models, using Llama-3-70B was argued to be 4.1 to 4400 times more carbon-efficient than a US human worker for simple writing tasks [31]. However, these models contained a critical methodological flaw: they did not control for output quality or correctness. They assumed that a single automated prompt yields a flawless, finalized document, whereas human professionals consume calories over several hours of planning, writing, and editing. Woo [33] addressed this quality-control issue by testing human programmers against GPT-4 models on complex programming tasks from the USA Computing Olympiad, requiring the AI models to execute iterative prompting, testing, and debugging loops until they produced a functionally correct code solution that matched the human programmers' output. Once output quality was controlled for, the efficiency narrative reversed: GPT-4 emitted 5 to 19 times more CO_2eq than human programmers to complete identical tasks under the specific experimental conditions reported by Woo [33], which involved competitive programming tasks from the USA Computing Olympiad, correctness-controlled iterative prompting, and amortized GPU manufacturing costs. These results should not be generalized to all programming tasks or educational settings without replication. In classroom settings, students rarely obtain perfect answers on their first prompt. They run multiple debugging, editing, and revision loops [20]. This corrected evidence demonstrates that AI-assisted education tasks can be significantly more carbon-intensive than traditional human execution once functional correctness is required.

4.5.2. Methodological Comparison

Early comparative studies used speculative mathematical formulas to translate human metabolic calorie consumption and professional lifestyle profiles into modern carbon equivalents, directly contrasting these with raw, single-prompt model API emissions estimates [32,48]. Woo [33] introduced a correctness-controlled comparative paradigm. By automating code compilation loops, Woo tracked real API usage, token generation, and physical database execution times across multiple rounds of problem-solving. Woo also integrated amortized GPU hardware manufacturing costs (Scope 3) to build a rigorous comparison of human work hours and AI computing hours.

4.5.3. Strength of Evidence

The corrected evidence showing that AI emissions exceed human carbon outputs for complex, multi-round tasks is supported by controlled experimental validation using standardized testing databases [33]. In contrast, early claims that AI has a massive carbon advantage remain highly speculative and methodologically flawed because they do not account for real-world academic failure rates, multi-turn debugging loops, or downstream reading and editing times.

4.5.4. Research Gaps

We found no comparative labor analyses evaluating subjects beyond competitive programming and simple descriptive writing. There is a complete lack of research comparing human and AI carbon costs in specialized tasks such as medical diagnostic training, qualitative research analysis, or graphic design. Research has also not examined how the carbon cost of teacher lesson planning changes when comparing manual preparation with multi-round generative prompting.

4.6. Theme 6: Educational Interventions, Student Awareness, and Behavioral Feedback

4.6.1. Synthesis

Survey evidence from international institutions shows that although students use GenAI daily, their environmental awareness remains superficial. General college students use AI tools for productivity without recognizing upstream water consumption, rare metal mining, or e-waste [38]. Undergraduate engineering students in India similarly lacked explicit knowledge of digital resource metrics, though they had an intuitive sense that data servers require substantial electricity [41]. To address this awareness gap, studies have evaluated direct interface interventions. Andersen et al. [18] added a real-time "CO_2 Tracker" to student writing platforms, showing that visual carbon feedback led students to voluntarily simplify queries, shorten prompt tokens, change models, and lower emissions (mean prompt token change: -8378, $p = 0.014$; CO_2 change: -1846 mg, $p = 0.009$). Prompt optimization can also be achieved structurally: Jung et al. [49] reduced student prompt token counts by 20% by translating queries from Korean to English and removing unnecessary words, demonstrating how language translation saves operational energy. However, when exposed to concrete environmental data, tech-educated young adults exhibit defensive psychological patterns. Ansari and Sameer [39] found that after students were exposed to concrete GPT-3 training examples, their perception of AI shifted dramatically, moving from an intangible, almost "spiritual" cloud service that they did not link to any material or environmental cost to viewing it as a concrete environmental threat, thereby initiating what the authors termed a "discursive inversion of accountability". While students willingly accepted personal blame for traditional physical waste, they externalized digital emissions entirely, arguing that platform developers and server operators are solely responsible for mitigation. This defensive shift allowed students to maintain an eco-conscious self-identity while continuing to use resource-intensive generative software for daily tasks without changing their behaviors.

4.6.2. Methodological Comparison

Educational studies primarily rely on self-reported survey instruments using Likert scales to measure student attitudes, knowledge levels, and digital habits [38,41]. More rigorous studies use mixed-method and experimental designs, combining pre- and post-intervention surveys with platform activity logs [18,20,39]. For example, Andersen et al. [18] linked user platform logs directly with mathematical emission calculators, allowing regression modeling of actual behavioral changes after the introduction of visual feedback. Qualitative studies use thematic analysis and focus group transcripts to map students' social beliefs and psychological defense strategies when discussing climate responsibility [39,42].

4.6.3. Strength of Evidence

The evidence showing that basic student awareness of digital footprints is low is supported by multiple survey studies, validated by multiple independent survey studies in different countries [30,38]. The evidence that interface feedback triggers conservation behavior is moderate, supported by one high-quality pilot trial [18], but requires validation in larger, more diverse student cohorts. The evidence documenting the psychological "discursive inversion of accountability" is moderate, with clear qualitative patterns observed in elite technical institutions, but broader testing is needed across non-technical domains.

4.6.4. Research Gaps

We found no research examining how digital footprint literacy or computer science sustainability modules are taught in primary or secondary schools. No empirical trials have evaluated digital sustainability interventions in adult education or professional training contexts. Furthermore, research has not investigated how teachers' attitudes toward environmental sustainability affect their integration of GenAI tools into classroom requirements.

4.7. Systematic Synthesis of Reviewed Evidence

Our systematic cross-study synthesis of the 23 included publications is summarized below, detailing the themes, contributing studies, level of empirical evidence, and dominant methodological approaches (Table 3). The accompanying evidence matrix (Table S3) provides the underlying study-level extraction data—including models, functional units, hardware assumptions, system boundaries, and uncertainty treatment—on which this cross-study synthesis is based.

Table 3. Cross-Study Empirical and Methodological Synthesis.

Theme	Studies Contributing Evidence	Level of Evidence	Main Methodological Approaches
Operational Electricity & Grid Demand	[11–13,15–17,20,45]	Strong (Hardware benchmarks & workspace trials)	Physical GPU Watt meters, on-site energy logging, bottom-up API software calculators, and PUE calculations.
Operational Carbon & GHG Footprints	[13,16,18,21,44]	Strong (LCA simulations & API logging trials)	Carbon-accounting packages (EcoLogits, Green Algorithms), grid emission intensity multipliers, API activity logging, and token count tracking.
Water Consumption & Cooling Demands	[15,16,39,43,44]	Limited (Restricted to secondary estimates)	Applying broad water-intensity coefficients to energy models, secondary corporate disclosures, and localized water-stress indexing.
Embodied Resources & e-Waste Lifecycles	[7,13,15–17,20,21,43]	Moderate (Strong technically, absent in education)	Multi-stage lifecycle assessment (LCA) software frameworks, industrial material database queries (Ecoinvent), and server replacement frequency models.
Comparative Labor Footprints	[32,33,48]	Strong (Controlled code-completion experiments)	Automated correctness compilation cycles, calorie-to-carbon conversions, software debugging models, and amortized manufacturing additions.
Educational Interventions & Student Dynamics	[18–20,38,39,41,42]	Moderate (Surveys & single-site pilot trials)	Self-reported Likert scale surveys, platform-embedded eco-feedback widgets, qualitative focus group transcript analyses, and prompt-level translation trials.

5. Discussion

Educational institutions at all levels—from primary and secondary schools to higher and adult education—are rapidly integrating generative artificial intelligence (GenAI) into lessons, homework, and administrative operations. However, this systematic review reveals a significant sustainability paradox. While educators and technology providers promote GenAI as a seamless, dematerialized tool, educational systems have largely overlooked its substantial physical resource footprint. Below, we examine the myth of weightless digital learning, investigate the psychological strategies users employ to avoid ecological guilt, clarify the technical calculations required to measure classroom emissions, and propose a practical policy framework for schools and universities.

5.1. The Myth of Dematerialized Digital Learning

For decades, educational technology has been promoted as a clean, sustainable alternative to physical classroom materials [50]. Transitioning from paper textbooks, handouts, and physical classrooms to web-based platforms was presented as an automatic environmental benefit. GenAI tools have reinforced this belief. Prompts are entered in a browser, and pages of text, custom lesson plans, or detailed graphics appear within seconds. There is no visible waste, making the interaction seem entirely weightless. The data gathered in this review challenge this illusion. As Lupetti et al. [20] found in their study of student designers, introducing generative AI tasks doubled the total energy footprint of students' educational activities. The frictionless nature of text prompts causes a severe rebound effect, also known as Jevons' Paradox [4]. When the human effort required to generate text, code, or artwork drops to nearly zero, the volume of generated material increases exponentially, creating a massive collective computational load across school networks. This operational load is only part of the problem. Technical reviews [16,17,21] show that GenAI relies on high-density GPU servers that consume enormous amounts of power and require large quantities of fresh water for server-room cooling [4,15]. They also drive significant "Scope 3" industrial impacts, such as toxic chemical waste and metal mining during hardware assembly, and accelerated e-waste cycles as servers are replaced every three to five years [7,16]. When a school or university encourages students to use commercial LLMs, it is not saving resources; it is shifting its resource footprint to distant data centers, cloud infrastructure networks, and mining regions, which are often located in vulnerable, water-stressed communities [4,15].

5.2. Inverted Accountability: How Users Handle Guilt

A key challenge for sustainability education is how learners and teachers respond to the environmental impact of their computing habits. Survey research shows that students feel a strong sense of responsibility for traditional environmental issues, such as recycling and turning off classroom lights, but have significant blind spots regarding digital technology [38,41]. When students learn about the actual resource consumption of AI servers, they display a defense mechanism that Ansari and Sameer [39] term the discursive inversion of accountability. While they readily accept personal responsibility for plastic waste or household energy use, they shift the blame for digital emissions to software companies and server farms, arguing that these entities are solely responsible for reducing energy use and purchasing green power. This mental shift allows users to maintain their eco-conscious identities while continuing to use high-impact software for schoolwork without guilt. Instead of moderating their usage, students tend to rely on "technological solutionism", assuming that corporate green pledges and improved algorithms will automatically resolve the issue [39]. This psychological barrier demonstrates why educational institutions cannot depend on simple informational campaigns; they must incorporate physical and architectural safeguards directly into their digital learning systems.

5.3. Closing the Methodological Divide

This systematic review identifies a significant methodological gap. Computer science papers measure physical GPU electricity consumption in watts under controlled database configurations [11], but educational researchers are limited to closed-source, commercially hosted APIs. As a result, education researchers must rely on simple proxies, primarily token counts [19].

https://doi.org/10.3390/su18147213

Although maintaining clean text and low token counts is a useful classroom practice [19], tokens do not correspond to energy use in a straightforward, linear manner. In reality, the energy footprint of a single prompt depends on a complex set of system variables:

- Model Architecture: Mixture-of-Experts (MoE) designs (such as DeepSeek V3) route questions to specific sub-modules, which can reduce active parameters and lower emissions by 90% compared to dense models [16].
- Hardware Efficiency: Power Usage Effectiveness (PUE) ratings of the target data center and GPU generation (NVIDIA H100 vs. older A100 architectures) dramatically change the power needed to process a token [11].
- Regional Grid Intensity: A query processed in a data center powered by coal-heavy electricity emits several times more CO_2eq than the same query sent to a hydro-powered server [13,14].

To make educational carbon metrics credible, researchers must use more sophisticated bottom-up estimation software. Tools such as EcoLogits [13] retrieve active hardware specifications and cloud API metrics directly, calculating emissions based on real-time server locations and grid intensities. Andersen et al. [18] showed that when this bottom-up estimation is integrated into a student platform as a real-time "CO_2 Tracker", voluntary changes in student behavior are substantial. Making invisible server calculations visible as feedback alters how students interact with software.

5.4. The Correctness-Control Fallacy in Human–AI Labor Studies

A major debate in the literature is whether AI work is cleaner than human intellectual work. Early comparative models [19,21] argued that AI writing and drawing offered a significant environmental advantage, claiming that computer prompts emitted 130 to 4400 times less carbon than human writers or artists. However, this early research suffered from a critical flaw: the correctness-control fallacy. These models assumed that a single, brief prompt would generate a perfect, final product on the first attempt. In actual schoolwork and writing, this is almost never the case. By controlling for correctness in coding tasks, Woo [33] reversed this narrative. When AI models were required to run iterative, multi-round prompts until they produced a functionally correct and verified solution, GPT-4 emitted 5 to 19 times more CO_2eq than a human programmer working on the same problem. This correction has serious implications for education. Learning is an iterative process. Students do not type one prompt to generate a final, perfect essay; they chat, debug, revise, and verify. When a student enters a dozen prompts to find a single correct answer, the cumulative resource cost of using large commercial LLMs increases rapidly [24]. These findings suggest that institutional policies attempting to replace human instruction, writing exercises, or tutoring tasks with GenAI systems under the banner of "eco-efficiency" should be approached with caution. However, the evidence base for this concern is currently limited to competitive programming tasks and simple writing comparisons. Whether replacing human educational activities with GenAI increases total emissions across different task types, quality standards, interaction rounds, and educational contexts remains an open empirical question that requires further verification. We therefore frame this as a research hypothesis warranting systematic investigation rather than a settled conclusion.

5.5. A Policy Framework for Sustainable Educational AI

To manage the carbon and material footprint of digital acceleration, schools, districts, and universities must move from short-term warning campaigns to concrete policy standards. We propose a three-tier framework presented below (Figure 2).

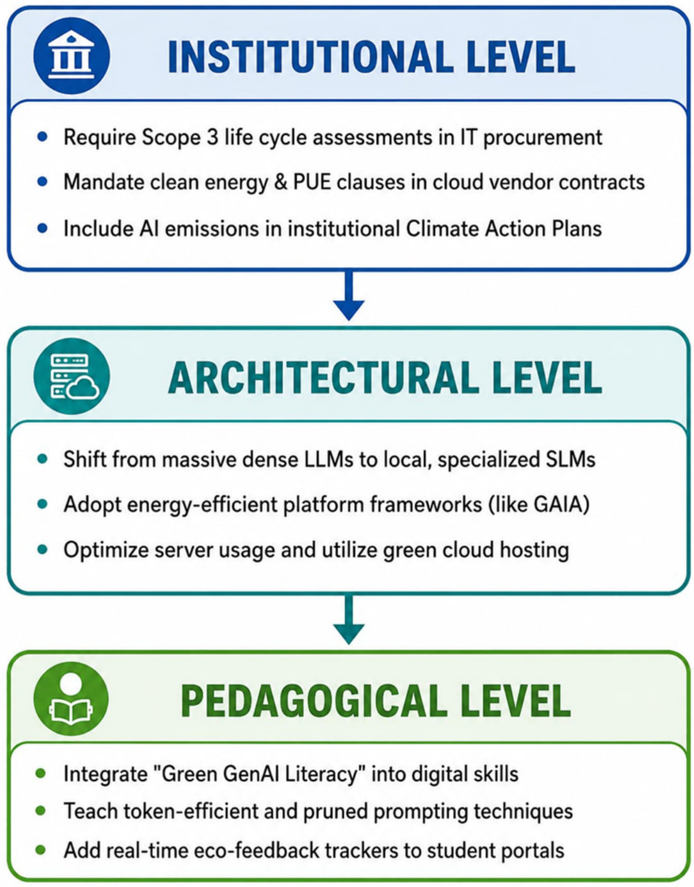

Figure 2. Core policy levels for sustainable AI adoption in education.

5.5.1. Classroom and Pedagogical Habits

Schools and universities should integrate “Green AI Literacy” into digital literacy workshops and curricula [4,40]. Immediate, low-cost modifications include:

- Token-saving techniques: Teaching learners prompt pruning and translation habits. As Jung et al. [19] showed, bilingual translation and concise drafting reduce token consumption by 6% to 20% without losing educational quality.
- Eco-feedback software: Adding visual trackers, such as the prompt carbon estimator developed by Andersen et al. [18], to school portals to help students develop more conscious prompting habits.

5.5.2. Technical and Platform Choices

IT administrators may wish to critically evaluate the default choice of massive, general-purpose cloud services against the specific needs of their educational context. The following options are presented as possible decision criteria for further investigation rather than as settled, evidence-based prescriptions.

- Specialized Small Language Models (SLMs): Multi-task models are orders of magnitude more power-hungry than specialized, task-specific models [51]. Schools and districts could consider running local, specialized SLMs (typically under 8 billion parameters) fine-tuned for educational tasks.
- Green local platforms: Building green architectures like GAIA [45] can cut platform energy consumption by 30% through optimized local server routing and lightweight model designs.

5.5.3. Institutional Governance and Purchasing

Educational administrators should use their purchasing budgets to demand transparency from software vendors.

- LCA procurement metrics: School board and university purchasing teams could request full Life Cycle Assessments (LCAs) for software systems, accounting for hardware manufacturing extraction, e-waste guidelines, and vendor carbon intensity [13,22].
- Green cloud agreements: Authorities could explore including environmental clauses in vendor contracts, with a goal of increasing the proportion of renewable energy powering educational server workloads over time, recognizing that immediate 100% locally additive renewable energy requirements may not be feasible for all institutions.
- Carbon auditing boundaries: Computing energy and data center emissions must be included in institutional Greenhouse Gas audits and Climate Action Plans (CAPs) [4]. This ensures that digital tools do not silently cancel out physical campus and building emission reductions.

6. Conclusions

This systematic review examined the environmental impacts of generative artificial intelligence in educational contexts by synthesizing evidence from 23 publications, including education-specific research and complementary technical studies. The findings reveal a substantial gap between the dematerialized experience of GenAI tools and their material environmental reality. The evidence base shows a sharp methodological divide. Technical studies use direct hardware measurements, lifecycle assessments, and carbon-accounting software that integrates regional grid intensity and model-specific benchmarks, producing precise estimates rarely applied in educational settings. Educational studies rely mainly on indirect proxies, with token counts serving as linear approximations for carbon emissions, despite significant variation introduced by model architecture and electricity sources. Survey instruments capture student attitudes rather than measured impacts. This divergence creates a paradox: the most rigorous environmental assessment methods are found in technical literature, while educational research lacks the tools and access to apply them directly.

Three findings stand out. First, the empirical evidence base in educational contexts is strikingly thin. Only three studies documented actual student prompting alongside environmental metrics—one showed that real-time carbon feedback reduced emissions, another demonstrated that GenAI doubled student energy consumption, and a third found that prompt pruning reduced token counts. All other educational publications rely on conceptual discussion or self-reported surveys. Second, comparative studies of AI versus human labor reveal a fundamental measurement problem: when studies control for output quality by requiring iterative prompting until functional correctness, AI-assisted tasks can emit 5 to 19 times more greenhouse gases than human workers. Third, student environmental awareness shows systematic blind spots, with learners recognizing traditional environmental issues but remaining unaware of digital resource demands. When presented with concrete data, students display inverted accountability, accepting responsibility for physical waste while externalizing responsibility for digital emissions to technology companies.

Critical gaps require further research. No studies have measured GenAI's environmental footprint in primary or secondary education. Longitudinal tracking of institutional-level adoption and aggregate impact is lacking. Water footprints and embodied emissions are discussed but not empirically measured in educational contexts. Comparative human–AI labor studies exist only for competitive programming and simple writing tasks, leaving translation, research synthesis, and administrative automation unexplored. To address these gaps, three interconnected interventions are needed. Institutions should integrate green AI literacy into curricula, teaching token-efficient prompting and meaningful interpretation of carbon feedback. Schools and universities should prioritize specialized small language models over general-purpose systems, implement carbon-aware infrastructure, and deploy real-time environmental tracking. Institutional procurement should require lifecycle assessment disclosure from vendors, include renewable energy requirements in cloud contracts, and incorporate GenAI emissions into greenhouse gas inventories.

The environmental impacts of generative AI in education are not incidental. They constitute a growing material footprint distributed across electricity grids, water systems, mineral supply chains, and waste streams worldwide. Current educational research and practice largely treat these impacts as external rather than integral to pedagogical and institutional decision-making. Closing this gap requires moving beyond awareness campaigns toward structural interventions that make environmental consequences visible, measurable, and actionable. The technical tools and methodological frameworks exist. The evidence base in educational settings must now expand to match the scale of adoption and the magnitude of environmental consequences at stake.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su18147213/s1, Table S1: Criteria; Table S2: Appraisal; Table S3: Evidence Matrix; Table S4. PRISMA_2020_checklist.

Author Contributions: Conceptualization, D.M.R. and T.K.; methodology, M.R.; formal analysis, M.R.; investigation, M.R.; resources, M.R., D.M.R. and T.K.; data curation, M.R.; writing—original draft preparation, D.M.R. and T.K.; writing—review and editing, M.R., D.M.R. and T.K.; visualization, M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Slovenian Research and Innovation Agency, grant number P5-0174 "Pedagogical-Andragogical Research—Learning and Education for Quality Community Life".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data extracted and analyzed in this review are provided in the Supplementary Materials.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT model 5.5 for the purposes of GenAI-assisted translation, editing of the text and image creation (Figure 2, the policy framework diagram). Figure 1 (the PRISMA flow diagram) was created by the authors without GenAI assistance. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADP	Abiotic Depletion Potential
API	Application Programming Interface
CAP	Climate Action Plan
GAIA	Green AI Analytics
GenAI	Generative Artificial Intelligence
GPU	Graphics Processing Unit
GPT	Generative Pre-trained Transformer
IEA	International Energy Agency
IT	Information Technology
LCA	Life Cycle Assessment

LLM	Large Language Model
MMAT	Mixed Methods Appraisal Tool
MoE	Mixture of Experts
NVML	NVIDIA System Management Interface
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PUE	Power Usage Effectiveness
SLM	Small Language Model

References

1. Konyrova, L. The evolution of language learning: Exploring ai's impact on teaching english as a second language. *Eurasian Sci. Rev.* **2024**, *2*, 133–138. [CrossRef]
2. Moybeka, A.M.S.; Syariatin, N.; Tatipang, D.P.; Mushthoza, D.A.; Dewi, N.P.J.L.; Tineh, S. Artificial Intelligence and English Classroom: The Implications of AI Toward EFL Students' Motivation. *Edumaspul-J. Pendidik.* **2023**, *7*, 2444–2454. [CrossRef]
3. Zhang, C.; Margulieux, L.E.; Haddadian, N. Integrating Generative Artificial Intelligence (GenAI) for Leveraging Early Literacy Instruction: A High-Leverage Teaching Practice (HLTP) Approach. In *Exploring the Potential of Artificial Intelligence in Early Childhood Education: New Learning Experiences for Young Children*; Routledge: New York, NY, USA, 2026; pp. 103–123. [CrossRef]
4. Bozkurt, A. AI's Thirst, AI's Heat, AI's Waste: Exposing the Hidden Environmental Impact of Every Artificial Intelligence Interaction. *Open Prax.* **2025**, *17*, 638–647. [CrossRef]
5. Pantelimon, F.-V.; Bologa, R.; Toma, A.; Posedaru, B.-S. The Evolution of AI-Driven Educational Systems during the COVID-19 Pandemic. *Sustainability* **2021**, *13*, 13501. [CrossRef]
6. van den Berg, G.; du Plessis, E. ChatGPT and Generative AI: Possibilities for Its Contribution to Lesson Planning, Critical Thinking and Openness in Teacher Education. *Educ. Sci.* **2023**, *13*, 998. [CrossRef]
7. Kneese, T.; Young, M. Carbon Emissions in the Tailpipe of Generative AI. *Harv. Data Sci. Rev.* **2024**. [CrossRef]
8. Lin, C.-C.; Huang, A.Y.Q.; Lu, O.H.T. Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review. *Smart Learn. Environ.* **2023**, *10*, 41. [CrossRef]
9. Morrison, J.; Na, C.; Fernandez, J.; Dettmers, T.; Strubell, E.; Dodge, J. Holistically Evaluating the Environmental Impact of Creating Language Models. *arXiv* **2025**, arXiv:2503.05804. [CrossRef]
10. IEA. *Energy and AI*; IEA: Paris, France, 2025.
11. Caravaca, F. Measuring Energy Consumption of LLMs Inferences. *Perform. Eval. Rev.* **2026**, *53*, 18–19. [CrossRef]
12. Luccioni, S.; Jernite, Y.; Strubell, E. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 85–99. [CrossRef]
13. Rincé, S.; Banse, A. EcoLogits: Evaluating the Environmental Impacts of Generative AI. *J. Open Source Softw.* **2025**, *10*, 7471. [CrossRef]
14. Bouza, L.; Bugeau, A.; Lannelongue, L. How to estimate carbon footprint when training deep learning models? A guide and review. *Environ. Res. Commun.* **2023**, *5*, 115014. [CrossRef] [PubMed]
15. Wu, S.; Xu, J.; Feng, S. Opportunities and Challenges of Generative AI in STEAM Education: A Thematic Co-occurrence Network Analysis of Primary and Secondary School Leaders' Perspectives. *Technol. Knowl. Learn.* **2026**. [CrossRef]
16. Esho, E.O.; Akinyelu, A.A.; Dinis, M.A.P. Sustainable generative AI and quantum computing: Review assessment on the environmental impact of generative AI and quantum technologies. *Front. Sustain.* **2026**, *7*, 1726832. [CrossRef]
17. Marmouzi, O.; Oumaira, I.; Ajana El Khaddar, M. A Systematic Review of Green and Sustainable AI: Taxonomy, Metrics, Challenges, and Open Research Directions. *Sustainability* **2026**, *18*, 4115. [CrossRef]
18. Andersen, L.B.; Herklotz, M.; Liu, A.; Goeke, M.; Juelich, M.; Kern, C.; Kreuter, F. From Awareness to Action? The Impact of CO_2 Emission Feedback on Student LLM Usage. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2026. [CrossRef]
19. Jung, H.; Min, J.; Moon, Y.; Son, J.; Whipple, M.E. Towards Sustainable Inference of LLMs for Medical Education through Token Count Minimization. *Stud. Health Technol. Inform.* **2025**, *329*, 1796–1797. [CrossRef] [PubMed]
20. Lupetti, M.L.; Cavallin, E.; Murray-Rust, D. The Unbearable Lightness of Prompting: A Critical Reflection on the Environmental Impact of genAI use in Design Education. *arXiv* **2025**, arXiv:2501.16061. [CrossRef]
21. d'Orgeval, A.; Sheehan, S.; Avenas, Q.; Assoumou, E.; Sessa, V. Generative AI impact assessment through a life cycle analysis of multiple data center typologies. *Appl. Energy* **2026**, *406*, 127288. [CrossRef]
22. Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; Pineau, J. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *J. Mach. Learn. Res.* **2020**, *21*, 248.
23. IEA. *Key Questions on Energy and AI*; IEA: Paris, France, 2026.

24. Kaack, L.H.; Donti, P.L.; Strubell, E.; Kamiya, G.; Creutzig, F.; Rolnick, D. Aligning artificial intelligence with climate change mitigation. *Nat. Clim. Change* **2022**, *12*, 518–527. [CrossRef]
25. Khajeh Naeeni, S.; Nouhi, N. The Environmental Impacts of AI and Digital Technologies. *AI Tech Behav. Soc. Sci.* **2023**, *1*, 11–18. [CrossRef]
26. Luccioli, V.; Iyengar, R.; Panley, R.; Haberkorn, F.; Ge, X.; Crane, L.; Sinha, N.; Lee, S.J. *LLM on a Budget: Active Knowledge Distillation for Efficient Classification of Large Text Corpora*; Board of Governors of the Federal Reserve System: Washington, DC, USA, 2025; pp. 2405–2414.
27. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3645–3650.
28. Mytton, D. Data center water consumption. *npj Clean Water* **2021**, *4*, 11. [CrossRef]
29. Wang, Y. Integrating large language models into medical undergraduate laboratory course to enhance bioethical competence: A quasi-experimental study. *Front. Med.* **2026**, *12*, 1745975. [CrossRef] [PubMed]
30. Cornelis, P.B.; Kuehr, R.; Yamamoto, T.; McDonald, R.; D'Angelo, E.; Althaf, S.; Bel, G.; Deubzer, O.; Fernandez-Cubillo, E.; Forti, V.; et al. *The Global E-Waste Monitor 2024*; International Telecommunication Union (ITU): Geneva, Switzerland; United Nations Institute for Training and Research (UNITAR): Bonn, Germany, 2024.
31. Ren, S.; Tomlinson, B.; Black, R.W.; Torrance, A.W. Reconciling the contrasting narratives on the environmental impact of large language models. *Sci. Rep.* **2024**, *14*, 26310. [CrossRef] [PubMed]
32. Tomlinson, B.; Black, R.W.; Patterson, D.J.; Torrance, A. The carbon emissions of writing and illustrating are lower for AI than for humans. *Sci. Rep.* **2023**, *14*, 3732. [CrossRef] [PubMed]
33. Woo, N.H. A comparative study of AI and human programming on environmental sustainability. *Sci. Rep.* **2025**, *15*, 39182. [CrossRef] [PubMed]
34. Rizzo, S.A. To be Artificial Intelligence for sustainability *or* not to be sustainable Artificial Intelligence. *Renew. Sustain. Energy Rev.* **2025**, *223*, 116063. [CrossRef]
35. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
36. Ouzzani, M.; Hammady, H.; Fedorowicz, Z.; Elmagarmid, A. Rayyan—A web and mobile app for systematic reviews. *Syst. Rev.* **2016**, *5*, 210. [CrossRef] [PubMed]
37. Hong, Q.N.; Fàbregues, S.; Bartlett, G.; Boardman, F.; Cargo, M.; Dagenais, P.; Gagnon, M.-P.; Griffiths, F.; Nicolau, B.; O'Cathain, A.; et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Educ. Inf.* **2018**, *34*, 285–291. [CrossRef]
38. Dungo, C.A.B.; Beltran, Z.L.E.; Declaro, B.C.; Dela-Cruz, J.J.C.; Viray, R.U. Students' Level of Awareness on the Environmental Implications of Generative AI. *J. Educ. Sci. Environ. Health* **2025**, *11*, 93–107. [CrossRef]
39. Ansari, B.; Sameer, A. The shifting locus of accountability: The discourse of young Indian adults (YIAs) on the environmental impact of artificial intelligence. *AI Soc.* **2026**. [CrossRef]
40. Wen, T.J.; Walker, K.; Xiao, C.J. Co-Designing Media Plans with Generative AI: Teaching Media Planning as Human–AI Collaboration in the Classroom. *J. Advert. Educ.* **2026**. [CrossRef]
41. Sivapragasam, C.; Vasudevan, M.; Natarajan, N.; Saravanan, P. Exploring the Challenges of LLMs in Higher Education: Is ChatGPT a Boon or Bane for the Students? *J. Eng. Educ. Transform.* **2026**, *39*, 574–583. [CrossRef]
42. Mbah, M.F.; Nugraha, T.R.; Kushnir, I. Challenges and Opportunities for Leveraging Generative AI for Sustainability Education: A Critical Review. *Sustainability* **2025**, *17*, 10623. [CrossRef]
43. Jetly, K.; Saif Al-Maqbali, M.; Nasser Said Al Fuliti, M. The Environmental Impact of Artificial Intelligence: Problems Possibilities and Solutions. *J. Phys. Conf. Ser.* **2026**, *3191*, 012056. [CrossRef]
44. Malik, F. Carbon-Conscious Intelligence: Life Cycle Assessment and Green Standards for Generative AI. *Int. J. Innov. Sci. Res. Technol.* **2025**, *10*, 1030–1037. [CrossRef]
45. Wen, X.; Lv, Y. Energy-efficient and secure digital learning using the GAIA green AI analytics framework. *Results Eng.* **2026**, *29*, 108746. [CrossRef]
46. *ISO 14040:2006*; Environmental Management—Life cycle assessment—Principles and Framework. International Organization for Standardization: Geneva, Switzerland, 2006.
47. *ISO 14044:2006*; Environmental Management—Life Cycle Assessment—Requireme. International Organization for Standardization: Geneva, Switzerland, 2006.
48. Ren, Y.; Wang, Y.; Dong, S.; Su, W.; Wu, Y.; Liang, S.; Yang, X. Application of Artificial Intelligence in Medical Education: A Systematic and Narrative Review of Pedagogical Potential and Ethical Implications. *Adv. Med. Educ. Pract.* **2026**, *17*, 567190. [CrossRef] [PubMed]

49. Jung, J.; Joo, S.; Park, H.; Han, I. Challenges of integrating ChatGPT into EFL writing: An Activity Theory framework. *Comput. Assist. Lang. Learn.* **2026**. [CrossRef]
50. Lait, J. Invisible energy policy and schools: How energy issues feature in the policies and documents of a UK secondary school. *Energy Effic.* **2023**, *16*, 58. [CrossRef]
51. Oliveira, L.; Tavares, C.; Strzelecki, A.; Silva, M. Prompting Minds: Evaluating How Students Perceive Generative AI's Critical Thinking Dispositions. *Electron. J. e-Learn.* **2025**, *23*, 1–18. [CrossRef]