CTE Workshop Proceedings, 2026, Vol. **13**, pp. 221–235 10.55056/cte.1357

Classification of artificial intelligence tools for educational research by the criterion of research autonomy

Tetiana A. Vakaliuk[1,2], Serhiy O. Semerikov[3,2], Oleh M. Spirin[2,4], Viacheslav V. Osadchyi[5,2] and Vasyl P. Oleksiuk[6,2]

[1]*Zhytomyr Polytechnic State University, 103 Chudnivsyka Str., Zhytomyr, 10005, Ukraine*
[2]*Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine*
[3]*Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*
[4]*Zhytomyr Ivan Franko State University, 30 Velyka Berdychivska Str., Zhytomyr, 10002, Ukraine*
[5]*Borys Grinchenko Kyiv Metropolitan University, 18/2 Bulvarno-Kudriavska Str., Kyiv, 04053, Ukraine*
[6]*Ternopil Volodymyr Hnatiuk National Pedagogical University, 2 Maxyma Kryvonosa Str., Ternopil, 46027, Ukraine*

Abstract. Existing frameworks classify AI tools for academic research by data type or functional role, leaving unanswered the question that most directly concerns research integrity: how much of the cognitive labour constitutive of scientific inquiry has been transferred to an algorithm? This paper proposes a classification built on a single criterion – *research autonomy* – defined as the degree to which a researcher retains control over the cognitive operations of scientific knowledge production. Five functional clusters form a spectrum from maximum to minimum research autonomy: (I) computational data analysis, where the algorithm performs only mathematically specified procedures; (II) content and discourse analysis, where it applies pre-validated category systems; (III) search and navigation, where it independently determines relevance; (IV) multimodal analysis, where it performs primary categorisation of pedagogical events; and (V) content generation and synthesis, where it generates text and proposes conceptual connections. For each cluster, the paper specifies educational research applications, characteristic methodological constraints, and ethical requirements. The framework supports three practical ends: methods reporting standards, cluster-differentiated institutional AI governance, and AI literacy curricula grounded in epistemic consequences.

Keywords: artificial intelligence, tool classification, educational research, research autonomy, academic integrity, research methodology

1. Introduction

Since 2022, the adoption of artificial intelligence (AI) tools in academic research has outpaced the development of methodological frameworks for evaluating their use [25]. Researchers increasingly rely on algorithms not merely for computation but for tasks that have traditionally constituted the cognitive core of scientific inquiry: identifying relevant literature, categorising discourse, recognising patterns in multimodal data, and synthesising heterogeneous information into conceptual frameworks. This shift raises questions that extend well beyond technical competence [9].

Existing attempts to classify AI tools for academic use have relied primarily on two approaches: grouping by the type of data processed, or by the functional role the tool fulfils [27]. Both approaches are practically convenient but methodologically incomplete. They do not answer the question that

0000-0001-6825-4697 (T. A. Vakaliuk); 0000-0003-0789-0272 (S. O. Semerikov); 0000-0002-9594-6602 (O. M. Spirin); 0000-0001-5659-4774 (V. V. Osadchyi); 0000-0003-2206-8447 (V. P. Oleksiuk)
vakaliuk@iitlt.gov.ua (T. A. Vakaliuk); semerikov@gmail.com (S. O. Semerikov); spirin@iitlt.gov.ua (O. M. Spirin); poliform55@gmail.com (V. V. Osadchyi); oleksyuk@iitlt.gov.ua (V. P. Oleksiuk)
https://acnsci.org/vakaliuk/ (T. A. Vakaliuk); https://acnsci.org/semerikov/ (S. O. Semerikov); https://nauka.gov.ua/researchers/rs.XeJkeyyH/ (O. M. Spirin); https://partner.kubg.edu.ua/contacts/rectorate/directors-and-deans/184-viacheslav-osadchyi.html (V. V. Osadchyi); https://tnpu.edu.ua/faculty/fizmat/oleksyuk-vasil-petrovich.php (V. P. Oleksiuk)

Received	Accepted	Published	Version of record
2026-02-18	2026-03-17	2026-03-21	2026-03-21

matters most to a researcher concerned with the integrity of their findings: *where in the research process does the boundary between human and machine agency lie, and what happens to the methodological status of results when that boundary shifts?* Two tools that both process text may have fundamentally different consequences – one counts lexical units according to criteria the researcher specifies; the other generates interpretive categories that the researcher can accept or reject but did not author.

Research in educational contexts is especially sensitive to this question. Educational research studies the processes of learning and cognition – precisely the processes that AI imitates or models – and its methodological toolkit includes qualitative approaches in which the researcher's subjective position is not a source of error but a constitutive component of the inquiry [11]. The absence of a systematic framework connecting tool choice to epistemic consequence leaves educational researchers without principled guidance for the most consequential decisions they face when using AI: not how to operate a tool, but what they are surrendering when they do.

We address this gap by proposing a classification of AI tools for educational research organised around a single criterion: *research autonomy* – the degree to which the researcher retains control over the cognitive operations constitutive of scientific knowledge production. The classification covers tools across all phases of the research process and is domain-specific: it concerns AI used *by educational researchers in their own research practice*, not AI deployed as a subject of study or as a teaching technology.

Two research questions guide the work:

RQ1: How can AI tools used in educational research be systematically classified by the degree of research autonomy they preserve?

RQ2: How can the methodological and ethical obligations imposed by each autonomy level be operationalised as a reporting standard applicable to editorial policy and researcher training?

Section 2 reviews related work. Sections 3–4 develop the theoretical framework and methodology. Section 5 presents the classification. Section 6 discusses implications and positions the framework against recent parallel proposals. Sections 7–8 address limitations and conclusions.

2. Related work

2.1. AI Tools in educational research

The systematic review by Zawacki-Richter et al. [27], covering publications from 2007 to 2018, identified four primary domains in which AI was applied in higher education: learning analytics and student profiling, intelligent tutoring systems, assessment, and adaptive systems. This four-domain taxonomy remains a useful landmark for understanding the field, though the authors themselves noted the scarcity of work examining AI as a tool *used by researchers* rather than as an object of study or an intervention in teaching. The arrival of generative language models after 2022 added a de facto fifth domain – support for academic writing and information synthesis – that the 2019 review did not anticipate [25].

UNESCO's guidance for generative AI in education and research [16] provided the first major institutional framework for the post-2022 landscape, identifying that generative models introduce research integrity risks that existing plagiarism-detection approaches do not address. Sperling et al. [24], in a scoping review of AI literacy in teacher education, found that most programmes respond to AI proliferation instrumentally – teaching how to use specific tools – without equipping researchers to reason about the epistemic consequences of delegating cognitive operations to algorithms.

Baker and Hawn [1], analysing automated assessment and learning analytics systems, demonstrated that these systems reproduce racial and linguistic inequalities when bias is not explicitly accounted for in the design and application of the tool. This finding extends beyond pedagogical practice to the research process itself: any tool that determines what is "relevant" in a dataset or body of literature embeds assumptions about what constitutes the norm.

2.2. Human-AI delegation and automation theory

The conceptual resources for understanding what happens when cognitive operations are transferred to technical artefacts have been developed in philosophy of technology since the 1980s. Callon [5] and Latour [13] established, within actor-network theory, that technical objects are not neutral intermediaries but carry embedded programmes of action – scripts that shape the behaviour of those who use them. In Latour's formulation, delegation is the displacement of human actions and intentions into material objects. Applied to AI tools in research, this delegation is not merely a matter of convenience or efficiency; it concerns the nature of what counts as knowledge when a significant portion of the cognitive labour is performed by an algorithm.

Molenaar [17] developed a six-level model of automation for hybrid human-AI learning environments, describing the distribution of control between person and algorithm from full human control through progressive stages of shared regulation to full machine control. Though designed for learning rather than research contexts, the structural logic of this model is directly applicable to the research process: the question of where the boundary of human agency lies is the same, and the consequences of mislocating it are analogous. The present paper transfers this logic from the learning to the research context, mapping Molenaar's six levels onto five functional clusters specific to educational research practice.

The sharpest recent empirical evidence for the importance of locating this boundary comes from Ríos-García et al. [20], who evaluated LLM-based scientific agents across eight domains in more than 25,000 agent runs. Their central finding is that these agents ignore available evidence in 68% of reasoning traces and do not exhibit refutation-driven belief revision – the epistemic pattern that makes scientific inquiry self-correcting. The base model accounts for 41.4% of explained variance in behaviour, while scaffold engineering accounts for only 1.5%. This finding provides direct empirical grounding for the present paper's treatment of Cluster V tools: an agent that generates text and proposes conceptual connections is not engaging in scientific reasoning, regardless of how the prompt is engineered.

2.3. Research integrity and authorship discourse

The normative dimension of AI use in research has been addressed most directly by the Committee on Publication Ethics, whose 2023 position statement [6] established that AI tools cannot be recognised as authors of scientific works and that their use must be disclosed. Bozkurt [3] frames this situation through the concept of co-creation, noting that responsibility for the content of a work cannot be transferred to an algorithm because responsibility presupposes the capacity for correction and accountability for claims – capacities that algorithms do not possess.

Koskinen [11] argues that the conceptual apparatus available to science studies is inadequate to describe research in which AI plays a significant role, and that the absence of an appropriate epistemology leaves researchers without principled guidance for evaluating the scientific status of AI-assisted findings.

Two recent frameworks address parts of this problem. Sanaei and Rajabzadeh [21] propose a two-dimensional framework for evaluating LLM applications in qualitative social science research, classifying tools along axes of interpretive depth and autonomy. Zhang et al. [28] develop a three-level taxonomy – Evaluator, Collaborator, and Scientist – for the roles that LLMs may play in scientific innovation more broadly. Both frameworks are discussed in section 6 in relation to the present proposal.

What the three streams share is a blind spot: no existing framework asks, simultaneously, which specific cognitive operations a tool displaces and to what degree the researcher retains control over them. Classification by data type and normative discourse about authorship address related concerns from opposite ends; a single criterion connecting delegation to epistemic consequence across all tool types and all phases of the research process is absent.

3. Theoretical framework

3.1. Delegation as transfer of epistemic agency

In actor-network theory, a technical object carries an embedded programme of action – a *script* that partly prescribes the behaviour of whoever uses it [5, 13]. Different scripts presuppose different competences and render different operations obsolete. Consider two contrasting cases. The script of SPSS running a multilevel regression presupposes that the researcher has specified the hypothesis, the model structure, and all variable definitions; it delegates only arithmetic. The script of a large language model asked to "explain the relationship between pedagogical context and student engagement" presupposes almost none of this: it generates the categorical connections itself, drawing on patterns in its training data rather than on the researcher's disciplinary knowledge. Delegation in the second case is not computational but *epistemic* – the tool decides what counts as a meaningful connection, a function that previously belonged to the researcher and that carries theoretical commitments shaping the status of the resulting knowledge claim.

3.2. The automation continuum as structural scaffold

Molenaar [17] models hybrid human-AI learning environments through six levels of automation, ranging from full human control (level 1) to full machine autonomy (level 6). Although developed for learning contexts, the structural question is the same for research: where does the boundary of human cognitive control lie, and what are the epistemic consequences when it shifts? The present paper maps Molenaar's six levels onto five functional clusters specific to the educational research process:

Levels 1–2 → Cluster I: the researcher specifies all analytical operations; the algorithm computes.

Level 3 → Cluster II: the researcher defines category systems; the algorithm classifies linguistic units.

Level 4 → Cluster III: the algorithm determines relevance; the researcher verifies.

Level 5 → Cluster IV: the algorithm performs primary categorisation of pedagogical events; the researcher interprets.

Level 6 → Cluster V: the algorithm generates text, arguments, and conceptual connections; the researcher accepts or rejects.

Molenaar's levels 1 and 2 are collapsed into a single cluster because the epistemic distinction between them – whether the system merely executes (level 1) or also suggests before executing (level 2) – does not alter the researcher's categorical authority in research contexts: in both cases, the researcher specifies every analytical operation and the algorithm performs only computation. Splitting them would create a distinction without a methodological difference for research practice [10].

3.3. Research autonomy: formal definition

The classification proposed in section 5 is organised around the following construct:

> *Research autonomy* is the degree to which a researcher retains control over the cognitive operations constitutive of scientific knowledge production: observation and classification of phenomena, detection of patterns in data, interpretation of pattern significance, synthesis of heterogeneous information into a conceptual frame, and generation of new concepts and theoretical propositions.

Lower research autonomy means more operations are delegated; it is a calibrative, not prescriptive, criterion – lower autonomy entails higher methodological obligations, not disqualification. Figure 1 illustrates the resulting spectrum.

Figure 1: The research autonomy spectrum: five functional clusters of AI tools for educational research, mapped to Molenaar's [17] automation levels. Each segment names the primary cognitive operation delegated to the algorithm.

3.4. On the dimensionality of research autonomy

Representing research autonomy as a single axis is a deliberate simplification. The five cognitive operations in the formal definition – observation and classification, pattern detection, interpretation, synthesis, and concept generation – are not independent: each presupposes the ones below it in the hierarchy. Concept generation presupposes synthesis; synthesis presupposes pattern detection. This hierarchical ordering justifies treating delegation as a matter of degree on a single scale rather than as combinations of independent dimensions. A multi-dimensional model remains a direction for future work [22].

Crucially, the unit of classification is the *practice*, not the tool in isolation. The same product may occupy different clusters depending on how it is used: GPT-4 operating as a structured classifier (researcher-specified categories, model applies them) functions at Cluster II; the same model asked to propose conceptual categories functions at Cluster V. This is addressed in the methodology (section 4) through the concept of dominant epistemic function.

Researchers can apply the following three diagnostic questions to place any AI-mediated practice on the spectrum:

1. *Who specifies the categories?* If the researcher defines all categories before the tool operates, the practice belongs to Cluster I or II. If the algorithm generates categories, it belongs to Cluster IV or V.

2. *Who determines which sources or data units are relevant?* If the researcher controls inclusion, the practice is Cluster I or II. If the algorithm determines visibility, it is Cluster III or higher.

3. *Who authors the conceptual claim?* If the researcher generates every argument and the tool only processes or formats it, the practice sits at Cluster I through IV. If the algorithm proposes the argument and the researcher evaluates it, the practice is Cluster V.

These questions do not replace the full cluster descriptions in section 5, but they provide a practical entry point for methods reporting and institutional governance decisions.

4. Methodology

The classification presented in section 5 was developed through *conceptual synthesis*: a theoretically grounded analytical procedure in which cluster boundaries are derived from the intersection of established automation theory and a systematic reading of the literature on AI tools in educational research. No formal review protocol was pre-registered; the approach follows Grant and Booth's [8] critical review typology, emphasising conceptual organisation over exhaustive coverage, combined with thematic synthesis in the sense of Braun and Clarke [4].

Scope. The classification covers AI tools used by educational researchers in their own research process: data collection and analysis, literature search and synthesis, and academic writing. Tools deployed as objects of study (e.g., AI tutors evaluated for learning outcomes) or as teaching technologies are outside scope.

Literature survey. We identified 47 candidate sources across the three bodies of literature reviewed in section 2. After excluding sources addressing AI as an object of study rather than a research instrument, 21 were retained for systematic analysis. Sources were drawn from peer-reviewed publications and authoritative institutional documents published between 2019 and 2026, identified through Semantic Scholar, Google Scholar, arXiv, and targeted citation tracing from three anchor papers: Zawacki-Richter et al. [27], and Molenaar [17].

Cluster derivation. Boundaries between clusters were determined by jointly applying two criteria: (a) the type of cognitive operation delegated, ranging from arithmetic computation to conceptual generation, and (b) the Molenaar automation level the tool's primary function most closely corresponds to. Molenaar's six levels map onto five clusters because levels 1 and 2 are epistemically equivalent in research contexts: in both cases the researcher retains full categorical authority and the algorithm performs only computation. Collapsing them avoids a distinction without a methodological difference (see section 3.2).

To illustrate how placement decisions were made: NVivo's AI autocoding belongs in Cluster I, not Cluster II, because the researcher specifies all thematic categories before the tool operates and the algorithm applies them to the data – matching Molenaar levels 1–2. LIWC-22 belongs in Cluster II because the categories are pre-validated dictionaries not defined by the researcher for the specific study, which corresponds to level 3. This distinction – researcher-specified categories versus pre-validated external systems – was the most frequently adjudicated boundary in the analysis. Cluster boundaries proved harder to draw than anticipated; the Cluster III–IV boundary in particular required adjudication across multiple tool examples before a consistent principle (locus of primary categorisation versus locus of relevance determination) emerged.

Tools straddling two clusters are placed by dominant epistemic function and flagged explicitly as boundary cases in section 5.

Positionality. The authors are affiliated with Ukrainian research institutions and developed the framework in a context of heightened concern for research integrity and AI governance. The framework's applicability in other regulatory environments has not been empirically tested, as noted in the limitations (section 7).

Validation. The framework was checked for internal consistency (each cluster is mutually exclusive on the delegation criterion), alignment with Molenaar's [17] automation continuum, and differentiation from the parallel frameworks of Sanaei and Rajabzadeh [21] and Zhang et al. [28]. All co-authors reviewed the cluster assignments iteratively.

5. Classification of AI tools for educational research by research autonomy

5.1. Cluster I: Computational data analysis (maximum research autonomy)

Cluster I represents the baseline against which the other clusters should be understood, not because computational tools are simple or epistemically undemanding, but because they leave the full architecture of the research intact. Every analytical decision – which hypothesis to test, which method to use, which parameters to specify – remains with the researcher; the algorithm performs only what has been completely specified. The result is that errors in Cluster I outputs are traceable to researcher decisions, not to algorithmic judgment. This traceability is what makes such tools compatible with the standards of scientific accountability without additional precautions beyond standard pre-registration and audit trail practices (Molenaar levels 1–2; see section 3.2).

Representative tools. Quantitative analysis is served by IBM SPSS AI modules (automatic model selection, anomaly detection), R with the tidymodels, caret, lavaan, and PyMC libraries, and the Python ecosystem of scikit-learn and statsmodels. For qualitative analysis, AI-assisted coding platforms – Atlas.ti, NVivo, and MAXQDA – belong to this cluster when used correctly, though they occupy a distinct sub-level: the algorithm proposes semantic similarity groupings, but the decision as to whether any proposed grouping is pedagogically meaningful rests entirely with the researcher [18].

Modern AI modules in these platforms detect thematic proximity, suggest variable transformations, and flag multicollinearity – tasks that previously required substantial statistical expertise.

Applications in educational research. A representative case is a national-scale study of student mathematics achievement employing multilevel modelling in R to account for the nested structure of the data (students within classrooms within schools): every model specification, interaction term, and random-effects structure is the researcher's decision; R computes. Similar logic applies to psychometric validation through confirmatory factor analysis, intervention effectiveness studies with repeated-measures designs, and large-scale comparative analyses that would be physically impossible to process by hand.

Methodological requirements. Pre-registration of hypotheses and the analysis plan before data collection prevents the practice of hypothesising after results are known (HARKing) – a risk that automation accentuates by making it trivially easy to test dozens of variable combinations and present the significant result as a primary hypothesis [1]. For qualitative AI-assisted coding, the researcher must verify every code the system proposes, document their own coding logic and positional decisions in the methods section, and maintain a complete decision trail for potential audit.

5.2. Cluster II: Content and discourse analysis (high research autonomy)

Cluster II tools perform pre-classification of semantic categories according to rules that are explicit, documented, and validated independently of any particular study (Molenaar level 3, see section 3.2). The difference from Cluster I is the locus of categorical authority: here, the categories are not invented by the researcher for the specific study but drawn from externally developed systems – validated dictionaries and parsers – that the researcher selects and applies.

Representative tools. The defining instrument of this cluster is LIWC-22 (Linguistic Inquiry and Word Count), which quantifies psychological, cognitive, and social categories in natural text by comparing each word against validated dictionaries covering more than eighty categories [19]. Alongside it sit Coh-Metrix and TAACO (Tool for the Automatic Analysis of Cohesion) for syntactic complexity and cognitive load measurement, and syntactic parsers such as spaCy and Stanford CoreNLP. A boundary case is GPT-4 used with a structured, researcher-authored prompt as a classification instrument: when the prompt fully specifies the categories and the model applies them without generative freedom, the tool functions at level 3; when the prompt invites the model to propose its own categories, it shifts to level 6.

Applications in educational research. A concrete case is LIWC analysis of Ukrainian-language teacher reflection journals, where the researcher must first audit whether the English-developed dictionaries capture the relevant cognitive categories with validity in a post-Soviet professional context – a methodological decision the algorithm cannot make. Other applications include analysis of ideological patterns in pedagogical communication, systematic textbook discourse analysis at a scale inaccessible to manual reading, and quantifying higher- and lower-order cognitive demands in classroom teacher talk.

Methodological requirements. The researcher using LIWC-22 is obligated to critically assess whether its dictionary categories correspond to their theoretical framework and cultural context: the system's dictionaries were developed primarily on English-language texts and may carry culturally specific associations that are not valid in other linguistic contexts. Results of automatic linguistic analysis must be interpreted exclusively in combination with deep contextual knowledge of the corpus; the algorithm responds to surface lexical and syntactic features but cannot detect irony, politeness as a strategy of power, culturally conditioned euphemism, or implicit pedagogical meaning embedded in subtext. Presenting the numerical output of Cluster II tools without this contextual qualification risks producing methodologically invalid conclusions clothed in false quantitative precision.

5.3. Cluster III: Search and navigation in scientific space (moderate research autonomy)

When an educational researcher uses Semantic Scholar to orient in a literature, an algorithm is deciding what is relevant – and in doing so, it is determining which fraction of the scientific field becomes visible. That determination was traditionally a researcher's exclusive function: it required disciplinary knowledge to recognise relevance across terminological variation and historical distance. Cluster III tools perform this function algorithmically (Molenaar level 4, see section 3.2). They do not interpret findings or draw conclusions, but they gate the researcher's epistemic horizon in ways that are not always transparent.

Representative tools. Semantic Scholar, Elicit, ResearchRabbit, Connected Papers, and Scite retrieve by semantic content rather than keyword alone, identifying thematically adjacent work even without shared terminology. Scite [26] goes further by classifying citations as supporting, contrasting, or neutral, enabling rapid identification of contested findings. Research agents in Claude, Gemini, and Perplexity represent the highest delegation zone within this cluster: they execute multi-step searches, cross-reference sources, and produce synthesised reports. Their autonomy approaches Cluster IV because they not only retrieve but partially synthesise; no agent-produced synthesis can enter a paper without exhaustive human verification [20].

Methodological requirements. For researchers working on Ukrainian-language or non-Anglophone educational contexts, the algorithmic bias of Cluster III tools deserves explicit attention: systematic under-indexing of non-English journals and atypical theoretical frameworks can produce a literature review that reflects primarily English-language research while presenting itself as a universal picture [1]. These biases must be named explicitly in the methods section; systematic review protocols following PRISMA [23] require manual verification and supplementary database searches to compensate.

5.4. Cluster IV: Multimodal analysis (limited research autonomy)

Consider a researcher with sixty hours of classroom video footage. Even with domain expertise and a well-developed coding scheme, manual transcription, segmentation, and annotation of that corpus is a months- long undertaking. Cluster IV tools automate these operations – transcribing, segmenting, and categorising pedagogical events from non-textual signals (Molenaar level 5, see section 3.2). Unlike the previous clusters, the algorithm here does not only process: it performs primary categorisation, determining which moment constitutes a teacher question, a student engagement event, or a turn-taking act. This is the most methodologically nascent cluster, with tool capabilities evolving faster than established validation frameworks.

Representative tools. Video analysis platforms including Vosaic, Annoto, and Kaltura AI automate transcription, segmentation, and annotation of recorded pedagogical interactions [15]. For detailed interaction analysis grounded in conversation-analytic traditions, ELAN and Transana remain the standard instruments for coding mathematical problem-solving discourse (following Krummheuer's [12] tradition) and other fine-grained sequential data. Multimodal language models – GPT-4V, Claude Vision, Gemini 1.5 Pro – extend automated interpretation to visual data: textbook illustrations, spatial classroom organisation, student artefacts.

Boundary case. GPT-4V used to transcribe video belongs in Cluster IV (signal processing and segmentation). If the same model is then prompted to propose interpretive categories for what it transcribed – without the researcher specifying those categories in advance – the practice crosses into Cluster V, because category generation becomes the algorithm's function.

Applications in educational research. Cluster IV tools enable systematic study of non-verbal aspects of pedagogical interaction (eye contact, teacher proxemic behaviour, attention patterns) with detail and reproducibility inaccessible through manual observation forms. Automated segmentation supports identification of recurring behavioural patterns across large observational datasets; multimodal analysis of learning materials enables corpus-scale curriculum audits for gender, racial, or cultural bias in textbook illustrations.

Methodological requirements and ethical obligations. Data sovereignty and third-party processing risks are present across all five clusters wherever data is transmitted to external services – researchers should verify data processing agreements at every cluster. Cluster IV introduces two additional obligations absent from earlier clusters. First, video data of pedagogical interactions captures recognisable images of minors, requiring written informed consent from all participants and guardians, detailed data management documentation, encrypted storage with strictly limited access, and specified destruction timelines – all reflected in an ethics committee-approved protocol. Second, automatic emotion recognition, which increasingly appears in video analytics platforms, cannot reliably infer emotional states from facial expressions alone and shows differential performance by skin tone and demographic composition of training data [2, 14]. UNESCO guidance explicitly discourages its use in educational contexts [16]. Including emotion recognition outputs in research conclusions without accounting for these limitations risks producing incorrect findings that reproduce the biases embedded in the algorithm.

5.5. Cluster V: Content generation and synthesis (minimum research autonomy)

The central distinction Cluster V forces a researcher to articulate is between *form* and *content*. Form – syntax, style, grammatical correctness, structural organisation – can be improved with language models without material violation of academic integrity, provided this is disclosed transparently as required by COPE [6]. Content – theoretical arguments, data interpretations, conceptual positions, scientific conclusions – must remain the researcher's own: delegating it to an algorithm severs the connection between the named signatory of the work and its actual intellectual subject. The distinction is harder to maintain in practice than in principle. A passage edited for clarity while preserving the researcher's argument differs fundamentally from a passage generated by the model in response to "explain the relationship between these two concepts" – even if the researcher agrees with the result.

Cluster V tools – Claude, ChatGPT, Gemini Advanced, and related generative language models – generate text, propose arguments, and formulate conceptual connections at Molenaar level 6 (see section 3.2). Research autonomy is minimal: the algorithm authors intellectual content that the researcher can accept, reject, or modify, but did not originate. Large language models generate text through statistical recombination of patterns in training data rather than through theory-driven hypothesis generation [20]. Research whose conceptual framework is formed primarily by model output rather than by the researcher's own engagement with primary sources may constrain originality, since the model's outputs tend toward what has already accumulated mass in the training corpus [28].

NotebookLM: a bounded-corpus variant. NotebookLM occupies a distinct position within this cluster because its synthesis capacity is bounded by documents the researcher uploads, not by an open training corpus. When a researcher loads fifty curriculum policy documents they have personally selected and verified, NotebookLM's synthesis operates on a corpus whose scope and quality the researcher controls. This substantially reduces hallucination risk and increases the verifiability of proposed connections: any claim can be traced to a specific document in the upload set. The tool operates more like a Cluster III semantic search over a controlled corpus than like an unconstrained generative system – a reminder that research autonomy is a practice-level criterion: the same tool family can support qualitatively different epistemic relationships depending on how it is used. Employing NotebookLM to identify non-obvious connections among already-read primary sources is methodologically sound; using it to substitute for reading those sources is not.

Conditions for responsible generative AI use. The preceding analysis should not be read as a blanket restriction on Cluster V tools. Several uses are compatible with research integrity [10]: (1) *Argument stress-testing*: the researcher presents a developed argument to the model and asks it to identify weaknesses or generate counterarguments; the researcher evaluates and responds. (2) *Form improvement*: the researcher supplies the completed argument and uses the model to improve clarity or reduce redundancy, retaining all content decisions. (3) *Hypothesis space expansion*: the model generates candidate research questions from a domain description; the researcher evaluates which,

if any, merit pursuit. In each case, the researcher retains authorship of the intellectual content; the model assists a bounded, verifiable subtask.[1]

The classification proposed above organizes five functional clusters in descending order of research autonomy, as illustrated in table 1.

Table 1
Summary classification matrix: AI tools for educational research classified by research autonomy criterion.

Cluster	AI role	Representative tools	Educational research applications	Key methodological requirement
Cluster I. Computational data analysis (maximum autonomy)	Computational intermediary; no interpretive role	SPSS AI, R/tidymodels, scikit-learn, Atlas.ti AI, NVivo AI	Large-scale studies; multilevel modelling; psychometric validation	Pre-registration; human interprets outputs; full audit trail
Cluster II. Content and discourse analysis (high autonomy)	Classifies linguistic units per pre-specified dictionaries	LIWC-22, CohMetrix, TAACO, spaCy, Stanford CoreNLP	Pedagogical discourse; cognitive load; curriculum audit	Cultural validity audit; contextual reading; domain knowledge required
Cluster III. Search and navigation (moderate autonomy)	Determines relevance; structures scientific field	Semantic Scholar, Elicit, ResearchRabbit, Scite, Connected Papers	Systematic reviews; citation network analysis; PRISMA screening	Verify sources manually; name English-language bias; agents as scaffolding only
Cluster IV. Multimodal analysis (limited autonomy)	Primary categorisation of pedagogical events from non-textual signal	Vosaic, Annoto, Kaltura AI, GPT-4V, Claude Vision, Gemini Vision	Classroom interaction; non-verbal behaviour; curriculum audit	Informed consent; ethics approval; no auto emotion recognition
Cluster V. Content generation and synthesis (minimum autonomy)	Generates text, proposes arguments, formulates conceptual connections	Claude, ChatGPT, Gemini Advanced, NotebookLM, Writefull	Writing support (form); hypothesis brainstorming; synthesis from own corpus	Disclose per COPE [6]; form editable, content not; verify all claims; NotebookLM on own corpus preferred

6. Discussion

6.1. Positioning against parallel frameworks

Table 2 positions this framework against two recent parallel proposals. Sanaei and Rajabzadeh [21] classify LLMs for qualitative social science along two axes (interpretive depth, autonomy), while Zhang et al. [28] develop a three-level system-centred taxonomy (Evaluator, Collaborator, Scientist) for LLM roles in scientific innovation generally. The present framework differs in scope (all AI tool types, all research phases), orientation (researcher-centred: what the researcher surrenders, not what the system does), and domain specificity (educational research, with cluster content reflecting its methodological traditions and ethical requirements).

The classification is descriptive in its five-cluster structure (it maps existing epistemic practices) and prescriptive in its normative layer (it specifies what methodological obligations follow from each cluster). Both aspects are intended.

[1]The authors used a Cluster V tool (Claude Sonnet 4.6, Anthropic) during the preparation of this manuscript for drafting assistance and TikZ figure generation. This use instantiates the boundary the framework identifies. All conceptual claims, cluster assignments, and normative arguments were generated and verified independently by the authors; AI-generated text was revised substantially. The AI Declaration at the end of this paper provides further detail.

Table 2
Comparison of three AI-in-research frameworks across six dimensions.

Dimension	This paper	Sanaei and Rajabzadeh [21]	Zhang et al. [28]
Unit of analysis	AI-mediated research practice	LLM application in qualitative study	LLM role in scientific process
Scope	All AI tool types; all research phases	LLMs only; qualitative phase	LLMs; scientific innovation broadly
Normative orientation	Researcher-centred (what is surrendered)	System-capability-centred	System-capability-centred
Domain specificity	Educational research	Social science generally	Science broadly
Granularity	5 clusters, single axis	2 axes, continuous	3 levels, discrete
Actionable output	Reporting standard, policy scaffold, curriculum	Tool selection guidance	Innovation stage assessment

6.2. Implications

For journal editors. Methods sections increasingly contain the sentence "AI was used for analysis" – no more informative than "a computer was used". Editors could require that methods sections specify the cluster, the cognitive operation delegated, and the verification procedure. This is not new bureaucracy but a restoration of the methodological transparency already required of non-AI methods [7]. Importantly, lower research autonomy is not a disqualification: NotebookLM applied to a carefully curated researcher corpus (Cluster V) may be methodologically sounder than an unverified Semantic Scholar export (Cluster III). The framework enables calibrated editorial judgment, not uniform restriction.

For university policy-makers. Institutional AI policies often conflate tool types from different clusters, producing either false permissiveness (permitting Cluster V authorship violations under the same umbrella as Cluster I computation) or false restriction (prohibiting Cluster I tools on the grounds of risks specific to Cluster V). The five-cluster structure provides a principled basis for differentiated governance calibrated to actual epistemic consequences. A cross-cutting consideration applies at Clusters II–V: wherever data is transmitted to external services, researchers must verify data processing agreements. What distinguishes Cluster IV is the additional requirement for informed consent and ethics committee approval due to biometric data; what distinguishes Cluster V is the authorship obligation.

For researcher training. Most current AI literacy programmes respond instrumentally by teaching tool operation [24]. The five-cluster framework offers an alternative scaffold organised around the three diagnostic questions from section 3.4: who specifies the categories, who determines relevance, who authors the conceptual claim. These questions give researchers a durable decision procedure that survives the inevitable obsolescence of any specific product list.

6.3. A reporting standard for methods sections

The practical output of the framework is a reporting standard requiring three elements per AI tool used: (1) the cluster assignment, (2) the cognitive operation delegated, and (3) the verification procedure. Table 3 illustrates the difference.

The compliant declaration names the cluster, operation, and verification procedure – the three elements that allow an editor, reviewer, or future researcher to assess the epistemic status of the methods independently.

Table 3
Compliant and non-compliant AI methods reporting.

Type	Example declaration
Compliant	"*Literature search*: Elicit (Cluster III, Molenaar level 4). The algorithm determined relevance via semantic similarity; all results were screened by [author initials], who retained final inclusion authority. *Thematic analysis*: NVivo AI autocoding (Cluster I, levels 1–2); researcher specified all categories a priori and verified each auto-suggested code against raw data. *Generative AI*: not used for conceptual or argumentative content."
Non-compliant	"AI tools were used to assist with the research."

7. Limitations and future research

Five limitations warrant acknowledgment:

1. *Conceptual, not empirical, derivation.* Cluster boundaries are analytically derived from the intersection of automation theory and the AI tools literature. They have not been validated through surveys of researcher behaviour or observational studies of tool use. The five-cluster count reflects a judgment about the minimum number that produces internally coherent and practically distinct groupings; a different number is possible with different criteria.

2. *Authorial positionality.* The framework was developed by researchers affiliated with Ukrainian institutions in a context of heightened concern for research integrity following Russia's full-scale invasion. Its applicability in other regulatory environments – the EU AI Act framework, US institutional review contexts, low-resource settings without consistent tool access – has not been empirically tested.

3. *Tool volatility.* Products cross cluster boundaries as capabilities evolve. NVivo's AI coding has become progressively more interpretively autonomous with each version. The framework therefore classifies cognitive functions, not specific products; researchers should assess a tool's current capability against the diagnostic questions in section 3.4 rather than relying on any static product list.

4. *Individual versus team research.* The framework assumes a single researcher making autonomous decisions. Team-based research with distributed cognitive labour – where hypothesis generation, coding, and synthesis are assigned to different members – involves a different configuration of autonomy that the current framework does not address.

5. *Environmental and access costs.* The methodological obligations identified for Cluster IV–V tools do not incorporate the environmental costs (energy, carbon) of inference at scale, nor do they address the unequal access to proprietary Cluster III–V tools across institutions and countries. These gaps are relevant to equity-oriented governance and constitute a direction for future work.

Future research directions follow from these limitations: (1) empirical validation through a mixed-methods study of researchers' tool use and autonomy perceptions across the five clusters; (2) longitudinal tracking of how specific products migrate across clusters as capabilities evolve; (3) examination of whether the five-cluster structure holds uniformly across STEM and humanities educational research, where epistemological commitments differ substantially.

8. Conclusions

What this paper enables that prior frameworks do not is a connection between delegation and methodological obligation within a single criterion applicable to all AI tool types and all phases of educational research. The five-cluster spectrum (section 5) and the diagnostic rubric (section 3.4) together give researchers, editors, and policy-makers a shared vocabulary: not whether AI was used, but *what was given up* when it was, and what that requires in methodological terms.

RQ1 is answered by the five-cluster spectrum structured on Molenaar's automation continuum [17]; the clusters and their defining epistemic operations are documented in full in sections 3 and 5 and summarised in table 1. We note that five clusters represents our best current judgment; the limitations (section 7) acknowledge openly that a different count is possible with different criteria, and we invite empirical challenge.

RQ2 is answered by the reporting standard proposed in section 6.3: methods sections should specify cluster membership, the cognitive operation delegated, and the verification procedure. The underlying logic is calibrative: lower research autonomy implies higher methodological obligations, not disqualification.

Three practical outputs follow. A *reporting standard* (table 3) makes the epistemic status of AI-assisted methods transparent to readers and reviewers. A *policy scaffold* enables institutional governance differentiated by the actual epistemic consequences at stake rather than by the name of the tool. A *curriculum scaffold* organises AI literacy education around the diagnostic questions in section 3.4 rather than around product tutorials that will be obsolete within a year.

The framework is offered as a stage-one contribution: conceptual and normative, awaiting empirical validation. The provisional character of cluster boundaries is a feature, not a defect – it invites the research community to test, refine, and, where necessary, revise.

Acknowledgments

The authors express sincere gratitude to Dr. Iryna S. Mintii for her invaluable assistance in the preparation of this work.

Funding

This research was funded by the National Research Foundation of Ukraine under grant No. 2025.07/0074 "Artificial Intelligence for Educational Research: Prediction, Modelling of Integration, and Digital Research Competencies" (competition "Advanced Science in Ukraine 2026–2028").

Declaration on Generative AI

During the preparation of this work the authors used AI-assisted tools in two of the five clusters identified by the framework herein.

Cluster III (Search / navigation). Semantic-search interfaces for Consensus, arXiv, PubMed, and bioRxiv were used to discover and screen candidate literature. Relevance judgements and final inclusion decisions were made exclusively by the authors.

Cluster V (Content generation / synthesis). Claude Sonnet 4.6 (Anthropic) was used to draft portions of the manuscript text, generate Ti*k*Z figure source code, and assist with synthesis across selected literature. All AI-generated content was critically reviewed and substantially revised by the authors; all conceptual claims, cluster assignments, and normative arguments were generated and verified independently by the authors.

No AI tools were used for data collection or empirical analysis. Authors affirm that the final text reflects their own understanding and judgement. The use of a Cluster V tool in a paper that argues for methodological caution with Cluster V tools is addressed reflexively in section 5, footnote 1: it

instantiates the practice boundary-case the framework identifies, and was managed through the verification protocol described in section 4.

References

[1] Baker, R.S. and Hawn, A., 2022. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, 32(4), pp.1052–1092. Available from: https://doi.org/10.1007/s40593-021-00285-9.

[2] Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M. and Pollak, S.D., 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, 20(1), pp.1–68. Available from: https://doi.org/10.1177/1529100619832930.

[3] Bozkurt, A., 2024. GenAI et al.: Cocreation, Authorship, Ownership, Academic Ethics and Integrity in a Time of Generative AI. *Open praxis*, 16(1), pp.1–10. Available from: https://doi.org/10.55982/openpraxis.16.1.654.

[4] Braun, V. and Clarke, V., 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), pp.77–101. Available from: https://doi.org/10.1191/1478088706qp063oa.

[5] Callon, M., 1984. Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review*, 32(S1), pp.196–233. Available from: https://doi.org/10.1111/j.1467-954X.1984.tb00113.x.

[6] COPE Council, 2024. COPE position - Authorship and AI - English. Available from: https://doi.org/10.24318/cCVRZBms.

[7] Frimpong, V., 2026. AI Disclosure Without Accountability: Paper Compliance and the Governance Limits of Transparency in Scientific Research. *Preprints*. Available from: https://doi.org/10.20944/preprints202604.0956.v1.

[8] Grant, M.J. and Booth, A., 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), pp.91–108. Available from: https://doi.org/10.1111/j.1471-1842.2009.00848.x.

[9] Holmes, W., Bialik, M. and Fadel, C., 2019. *Artificial Intelligence in Education: Promise and Implications for Teaching and Learning*. Boston, MA: Center for Curriculum Redesign. Available from: https://www.researchgate.net/publication/332180327.

[10] Holmes, W. and Porayska-Pomsta, K., eds, 2022. *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*. New York: Routledge. Available from: https://doi.org/10.4324/9780429329067.

[11] Koskinen, I., 2024. We Have No Satisfactory Social Epistemology of AI-Based Science. *Social Epistemology*, 38(4), pp.458–475. Available from: https://doi.org/10.1080/02691728.2023.2286253.

[12] Krummheuer, G., 2015. Methods for Reconstructing Processes of Argumentation and Participation in Primary Mathematics Classroom Interaction. In: A. Bikner-Ahsbahs, C. Knipping and N. Presmeg, eds. *Approaches to Qualitative Research in Mathematics Education*. Dordrecht: Springer, Advances in Mathematics Education, pp.51–74. Available from: https://doi.org/10.1007/978-94-017-9181-6_3.

[13] Latour, B., 1992. Where are the missing masses? The sociology of a few mundane artifacts. In: W. Bijker and J. Law, eds. *Shaping Technology/Building Society: Studies in Sociotechnical Change*. Cambridge, MA: MIT Press, pp.225–259. Available from: http://www.bruno-latour.fr/node/258.html.

[14] Li, S. and Deng, W., 2022. A Deeper Look at Facial Expression Dataset Bias. *IEEE Transactions on Affective Computing*, 13(2), pp.881–893. Available from: https://doi.org/10.1109/TAFFC.2020.2973158.

[15] McLean, L. and Connor, C.M., 2018. Challenges, benefits, and considerations when conducting classroom video observation research. *Sage Research Methods Cases Part 2*. SAGE Publications,

Ltd. Available from: https://doi.org/10.4135/9781526436252.
[16] Miao, F. and Holmes, W., 2023. *Guidance for generative AI in education and research*. Paris: UNESCO. Available from: https://doi.org/10.54675/EWZM9535.
[17] Molenaar, I., 2022. Towards hybrid human-AI learning technologies. *European Journal of Education*, 57(4), pp.632–645. Available from: https://doi.org/10.1111/ejed.12527.
[18] Mortelmans, D., 2025. NVivo and AI: (Semi)-Automatic Coding. *Doing Qualitative Data Analysis with NVivo*. Cham: Springer Nature Switzerland, Springer Texts in Social Sciences, pp.229–250. Available from: https://doi.org/10.1007/978-3-031-66014-6_19.
[19] Pennebaker, J.W., Boyd, R.L., Jordan, K. and Blackburn, K., 2015. *The Development and Psychometric Properties of LIWC2015*. Austin, TX: The University of Texas at Austin. https://www.researchgate.net/publication/282124505, Available from: https://doi.org/10.15781/T29G6Z.
[20] Ríos-García, M., Alampara, N., Gupta, C., Mandal, I., Mannan, S., Aghajani, A.A., Krishnan, N.M.A. and Jablonka, K.M., 2026. AI scientists produce results without reasoning scientifically. 2604.18805, Available from: https://doi.org/10.48550/arXiv.2604.18805.
[21] Sanaei, A. and Rajabzadeh, A., 2025. Depth and Autonomy: A Framework for Evaluating LLM Applications in Social Science Research. 2510.25432, Available from: https://doi.org/10.48550/arXiv.2510.25432.
[22] Selwyn, N., 2019. *Should Robots Replace Teachers? AI and the Future of Education*. Cambridge: Polity Press.
[23] Souifi, L., Khabou, N., Rodriguez, I. and Kacem, A., 2024. Towards the Use of AI-Based Tools for Systematic Literature Review. *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*. INSTICC, SciTePress, pp.595–603. Available from: https://doi.org/10.5220/0012467700003636.
[24] Sperling, K., Stenberg, C.J., McGrath, C., Åkerfeldt, A., Heintz, F. and Stenliden, L., 2024. In search of artificial intelligence (AI) literacy in teacher education: A scoping review. *Computers and Education Open*, 6, p.100169. Available from: https://doi.org/10.1016/j.caeo.2024.100169.
[25] Stokel-Walker, C. and Van Noorden, R., 2023. What ChatGPT and generative AI mean for science. *Nature*, 614(7947), pp.214–216. Available from: https://doi.org/10.1038/d41586-023-00340-6.
[26] Tay, A., 2024. Google Scholar vs other AI search tools (Undermind, Elicit, SciSpace) – how and when to use each. Available from: https://doi.org/10.59350/33f86-7qn88.
[27] Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F., 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), p.39. Available from: https://doi.org/10.1186/s41239-019-0171-0.
[28] Zhang, H., Li, R., Zhang, Y., Xiao, T., Chen, J., Ding, J. and Chen, H., 2025. The Evolving Role of Large Language Models in Scientific Innovation: Evaluator, Collaborator, and Scientist. 2507.11810, Available from: https://doi.org/10.48550/arXiv.2507.11810.