*CTE Workshop Proceedings*, 2026, Vol. **13**, pp. 236–255 10.55056/cte.1397

# Reporting AI in education research: a methodological audit of 2025–2026 publications against an adapted TRIPOD-LLM checklist

Iryna S. Mintii[1,2,3,4,5,6], Dmytro V. Verbovetskyi[1] and Ostap Yu. Sirenko[1]

[1]*Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine*
[2]*University of Łódź, 68 Gabriela Narutowicza Str., 90-136 Łódź, Poland*
[3]*Lviv Polytechnic National University, 12 Stepana Bandery Str., Lviv, 79000, Ukraine*
[4]*Zhytomyr Polytechnic State University, 103 Chudnivska Str., Zhytomyr, 10005, Ukraine*
[5]*Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*
[6]*Academy of Cognitive and Natural Sciences, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*

**Abstract.** We audit how the use of artificial intelligence is reported in recent education research. From a harvest of 29 848 arXiv preprints and 543 articles in six education and education-technology journals (2025–2026), we coded 220 papers (127 arXiv + 93 journal) against a 19-item checklist adapted from the TRIPOD-LLM reporting guideline, plus descriptive and outcome items. Coding was performed by open-weight large language models (served through Ollama) from titles and abstracts, conservatively (an item not stated is coded 0); we report cross-model agreement (mean Cohen's $\kappa = 0.53$, raw agreement 87%) in place of inter-human reliability, and disclose the AI-coded method in full. Overall reporting compliance is low: the median paper reports 32% of the checklist items, and the lowest-compliance items are the cross-cutting accountability signals – funding and conflicts of interest, missing-data handling, calibration/fairness, compute and cost, and the human-in-the-loop protocol (each $\leq 7\%$). Reporting *quantity* does not differ between arXiv preprints and journal articles in the unadjusted comparison (equal medians; unadjusted odds ratio $\approx 1$); what differs is *composition* – preprints document the model machinery while journal articles document the study context, and neither documents accountability. A modest journal advantage in quantity emerges only after adjusting for study design. Empirical design is the dominant predictor of how many items a paper reports. A within-paper preprint-vs-published comparison was planned but could not be conducted, as no eligible pairs exist. We contribute the *TRIPOD-LLM-for-education* checklist – to our knowledge the first reporting checklist derived from TRIPOD-LLM and calibrated for general (non-medical) education research – as a citable artefact, and call on education journals to require accountability reporting at submission.

**Keywords:** artificial intelligence, education research, reporting quality, TRIPOD-LLM, PRISMA, methodological audit, reproducibility, transparency

## 1. Introduction

The period from the public release of ChatGPT in November 2022 through early 2026 has seen a remarkable surge in the use of artificial intelligence in education research. Bibliometric analyses suggest that papers combining the words "AI" and "education" in their abstracts have grown at roughly three times the rate of education-research publications overall, with the steepest acceleration concentrated in the 2024–2025 window [10, 12]. The 2025–2026 publication window examined in this audit is therefore the most intense 15-month slice in the history of the field, and almost certainly the period in which the largest number of *reporting decisions* – model name, version, prompt, training

0000-0003-3586-4311 (I. S. Mintii); 0000-0002-4716-9968 (D. V. Verbovetskyi); 0009-0006-4489-2110 (O. Yu. Sirenko)
mintii@iitlt.gov.ua (I. S. Mintii); verbovetskyj.dv@iitlt.gov.ua (D. V. Verbovetskyi); sirenko.ostap@iitlt.gov.ua (O. Yu. Sirenko)
https://acnsci.org/mintii/ (I. S. Mintii); https://nauka.gov.ua/researchers/rs.QfTbbfyM/ (D. V. Verbovetskyi)

| Received | Accepted | Published | Version of record |
|---|---|---|---|
| 2026-02-01 | 2026-03-19 | 2026-03-21 | 2026-03-21 |

data, evaluation protocol, reproducibility artefacts – were taken under the highest time pressure. Whether those decisions were taken well is the empirical question of this paper.

At the same time, the social and life sciences have entered the second decade of a documented reproducibility crisis [7, 13]. In education research, the Open Science Collaboration's 2015 estimate that fewer than half of psychology studies could be replicated has been extended to intervention research, with several systematic replications producing smaller-than-expected effect sizes. Generative AI has sharpened this concern in two distinct ways. First, AI introduces a new class of *non-deterministic* research tool whose behaviour is sensitive to prompts, model versions, and sampling temperature; without precise reporting, replication becomes essentially impossible. Second, AI itself can now be used to generate, edit, or review scientific text, raising the question of how much of a given paper was produced by the named authors and how much by an opaque language model. The conversation about AI in education has therefore become entangled with the broader conversation about reporting quality, but until now no audit has measured that entanglement empirically.

This paper addresses that gap with a *methodological audit*[1]: a structured, PRISMA-2020-anchored review of 2025–2026 publications that use AI in education research, scored against a 19-item checklist adapted from the TRIPOD-LLM statement [5]. Our specific research questions are:

1. **RQ1 (Inventory).** What proportion of 2025–2026 publications using AI in education research (a) report the AI method with sufficient detail to reproduce, and (b) align with the items of the TRIPOD-LLM-for-education checklist (the adapted rubric we introduce in section 3)?
2. **RQ2 (Source differences).** Does reporting completeness differ by source channel, comparing arXiv preprints with peer-reviewed journals (*Computers and Education*, *British Journal of Educational Technology*, *Internet and Higher Education*, *Computers and Education: Artificial Intelligence*, *Educational Researcher*, and *Review of Educational Research*)?
3. **RQ3 (Predictive factors).** Which paper-level features – sample size, design type, AI role, declared funding, country of first author – predict compliance with the rubric?
4. **RQ4 (Within-paper comparison).** For papers that exist as both an arXiv preprint and a peer-reviewed published version, does reporting completeness change between the two? This paired design controls for paper identity and isolates the channel effect.

The contributions are threefold. *Empirically*, this is the first stratified audit of 2025–2026 AI-in-education reporting practices, providing direct evidence about which reporting gaps matter most. *Methodologically*, we adapt the 19-item TRIPOD-LLM checklist to education research via a five-expert panel and publish it as the *TRIPOD-LLM-for-education* checklist, a citable artefact that the community can debate, refine, or adopt. *Practically*, we release a reproducible audit pipeline and an open coded dataset as supplementary material, so the audit can be repeated annually and the field can monitor its own progress.

The paper is organised as follows. Section 2 situates the audit in the prior literature on AI in education research, on reporting guidelines (TRIPOD-LLM, DOME), and on the meta-research tradition of methodological audits. Section 3 describes the methods, including the PRISMA-2020 protocol, the eligibility criteria, the search strategy, the LLM coding scheme, and the cross-model reliability check. Section 4 reports the per-RQ results on the 220 coded papers. Section 5 discusses the implications, including a proposed TRIPOD-LLM-for-education checklist and a set of recommendations for journals. Section 6 acknowledges the limitations. Section 7 concludes with a concrete call to action and a one-shot replication protocol.

## 2. Background and prior literature

The audit reported here sits at the intersection of four literatures: a critical literature on AI in education, a reproducibility-and-reporting literature, a methodological-audit (meta-research) literature, and a more recent literature on reporting guidelines for AI-based research.

[1] **Note on terminology.** A *methodological audit* is a structured assessment of the *reporting practices* of a literature, not a synthesis of the literature's findings; it is closer to a meta-research survey than to a systematic review.

### 2.1. A critical view of AI in education

The earliest sustained critique of AI in education is by Holmes et al. [8], who argued that many of the field's claims about AI-driven personalisation rested on a thin empirical base and that the field was importing an instrumental view of pedagogy incompatible with human-centred teaching. The contribution is now widely cited and is often taken as a reference point for subsequent criticism. Crompton and Burke [6] extend the argument to generative AI, noting that the field's enthusiasm for tools such as ChatGPT was running ahead of any empirical evidence about learning gains and that the "AI-can-do-everything" narrative was crowding out more pedagogically-grounded work. Bearman, Ryan and Ajjawi [3] address the authorship question directly, arguing that chatbot-generated text should not count as authorship and that journals need explicit policies on AI use. These three works are cited frequently in the audit corpus (section 4) and frame the normative position from which the present audit proceeds.

### 2.2. Reporting guidelines for AI in research

The most prominent AI-specific reporting guidelines are TRIPOD-LLM [5] and DOME [16]. TRIPOD-LLM is a 2024 update of the original 2015 TRIPOD statement, expanding the 22-item checklist to 19 items covering prediction-model development, validation, and update using regression or machine-learning methods. It was developed by an international consensus group and has been adopted by several clinical journals. DOME (*Data Optimisation Model Evaluation*) is a similar checklist for machine-learning in life-sciences research, with a focus on dataset description, model architecture, and evaluation protocol. Both guidelines were produced for clinical and life-sciences contexts; neither is calibrated to the kinds of studies typical of education research, where designs are often smaller, less standardised, and less frequently pre-registered.

Several recent reporting guidelines are specific to education research. The RAISE framework [2] (*Reporting AI Studies in Education*) is a 30-item checklist across 10 thematic domains, published in the *Journal of Educational Computing Research* in 2026. AITEL (*Artificial Intelligence Technology Enhanced Learning*) [9] is a checklist for AI in medical-education research, published in *Medical Teacher* in 2024. Other recent contributions include the TITAN guideline [1] for general manuscript reporting, the METRICS checklist [15] for generative-AI studies in health-professions education, and AMEE Guide 192 [4] on disclosing generative-AI use in academic publishing. These guidelines overlap in scope with the present audit's TRIPOD-LLM-for-education checklist but differ in origin and emphasis: RAISE is a bottom-up framework built from a qualitative review of the AI-in-education literature; AITEL is scoped to medical education; TITAN, METRICS, and AMEE 192 are general. The TRIPOD-LLM-for-education checklist proposed in the present audit is, to our knowledge, the *first* to derive from TRIPOD-LLM specifically and to be calibrated for general (non-medical) education research. The three frameworks (RAISE, AITEL, and TRIPOD-LLM-for-education) are complementary, not redundant; the existence of RAISE strengthens the audit's case for a *TRIPOD-LLM-for-education* complement.

### 2.3. Systematic-review and meta-research methods

The audit's methodology draws on the PRISMA-2020 statement [14] and its protocol sibling [11], the most widely-cited reporting guidelines for systematic reviews and meta-analyses. PRISMA-2020 specifies a four-phase flow diagram (identification, screening, included) and a 27-item reporting checklist; we follow the diagram exactly and adapt the checklist to a methodological-audit context (data not extracted, so items on summary measures, synthesis methods, and certainty assessment do not apply). The audit also adapts the established meta-research practice of *independent double coding*: here the independent coders are three separate LLMs rather than two humans, and we report Cohen's $\kappa$ and raw agreement between models per item as a check on the coding instrument's stability (section 3.6).

### 2.4. Recent AI-in-education meta-research

A handful of recent meta-research papers have surveyed the AI in education literature. Most are bibliometric or scientometric analyses that characterise the field by output volume, country distribution, and topic clusters. To our knowledge, none has focused specifically on the *reporting* of AI methods, and none has applied a TRIPOD-derived rubric. The closest precedent is a 2024 scoping review of generative AI in higher education that included a brief quality appraisal but used a generic JBI checklist rather than a discipline-specific one. The audit reported here differs in three ways: (a) it is restricted to a defined 15-month window, allowing the inference to be sharp about the current state of the field rather than a multi-year average; (b) it stratifies the corpus by source channel, allowing direct comparison between arXiv preprints and peer-reviewed journals; and (c) it includes a within-paper paired analysis that, by holding paper identity constant, isolates the effect of peer review on reporting completeness.

## 3. Methods

The audit follows a written protocol that fixes the eligibility criteria, the search strategy, and the analysis plan in advance, in line with the PRISMA-2020 protocol-statement guidance [11, 14].

### 3.1. Design

We followed the PRISMA-2020 framework [14] for the identification, screening, eligibility, and inclusion steps, and we adapted the PRISMA-2020 27-item reporting checklist to the methodological-audit context (the data-extraction, synthesis, and certainty-assessment items do not apply). The eligibility criteria and the analysis plan were specified in the written protocol before coding, to limit post-hoc refinement.

### 3.2. Eligibility criteria

We included original empirical research or methodological work published in English between 1 January 2025 and 15 March 2026 that used AI (large language models, classical machine learning, deep learning, or generative AI) in an education context (K-12, higher education, lifelong learning, or professional development). Both peer-reviewed journal articles and arXiv preprints were eligible. Six target journals were searched: *Computers and Education* (Elsevier), *British Journal of Educational Technology* (Wiley), *Internet and Higher Education* (Elsevier), *Computers and Education: Artificial Intelligence* (Elsevier), *Educational Researcher* (SAGE), and *Review of Educational Research* (SAGE). The arXiv categories searched were `cs.CY`, `cs.AI`, and `cs.CL` with at least one of *education*, *learning*, *students*, *teaching* in the title or abstract. We excluded pure conceptual or theoretical essays without data, secondary reviews and meta-analyses, non-English texts, retracted preprints, and papers that used "AI" loosely (e.g., for spreadsheet macros, basic statistics, or generic search).

### 3.3. Information sources and search strategy

The arXiv corpus was retrieved through the OAI-PMH endpoint at https://export.arxiv.org/oai2, queried with the set names `cs:cs:CY`, `cs:cs:AI`, and `cs:cs:CL` (the standard arXiv OAI-PMH format for the `cs.CY`, `cs.AI`, and `cs.CL` categories) and saved as raw JSONL (29 848 records; see Appendix A.1 for the exact per-month query). The OAI-PMH harvest was the fallback after the regular arXiv API was rate-limited; the query strings, the date window, the page size, the set names, and the per-chunk record counts are all recorded in `data/protocol/oai_fetch.log` for full reproducibility. The journal corpus was retrieved through the Elsevier ScienceDirect Search API, the Wiley Online Library search API, and a fallback scrape of the SAGE table-of-contents pages for the two SAGE journals, which do not expose a public search API.

### 3.4. Selection process

Titles and abstracts were screened against the inclusion and exclusion criteria conservatively: a heuristic regex screen (`scripts/screen_titles_heuristic.py`, decision rules logged) removed clearly out-of-scope records, and the remaining records were carried into the coding step, where the LLM coder's free-text rationale flags any residual scope mismatch (e.g. a non-education AI paper that survived the keyword filter). We did not establish a human gold standard for screening, so no screening-validity $\kappa$ is reported; the inclusion decision is deliberately permissive, and out-of-scope papers simply receive 0 on the education-specific items rather than being discarded.

All 27 items were then coded by the LLM from each paper's title and abstract – there was no full-text retrieval pass. Items that typically require the full text (e.g. missing-data handling, calibration, compute, data/code availability) are therefore coded conservatively (0 when not stated in the abstract), which makes their compliance rates lower bounds; this abstract-vs-full-text split is set out in section 3.5 and the limitation is quantified in section 6.

### 3.5. Coding scheme

The codebook was designed by the adaptation panel (below), but the *operational coder is a large language model*, not a human rater: each paper's title and abstract were submitted to an open-weight model served through Ollama (`minimax-m3` as the primary coder), which returned a structured 27-item code with a free-text rationale. We made this choice deliberately – a study of how AI use is reported that is itself AI-coded, and says so – and we report its consequences (conservative, abstract-level codes; cross-model rather than inter-human reliability) transparently throughout. The coding scheme has three layers. The *descriptive layer* records five paper-level features: source channel, year, country of first author, AI role (object of study, research tool, or both), and empirical design type (RCT, quasi-experimental, survey, qualitative, mixed methods, simulation/benchmark). The *TRIPOD-LLM-adapted layer* records 19 items, each rated 0 (not reported), 1 (partially reported), or 2 (fully reported), with a free-text rationale. The adaptation was performed by a five-expert panel convened virtually, comprising two methodologists, one AI/ML researcher, one learning scientist, and one journal editor. The panel worked in two passes: an open discussion of the spirit of each TRIPOD-LLM item, followed by a 0/1/2 rating on three pilot papers. These 19 adapted items constitute the *TRIPOD-LLM-for-education* checklist, presented as table 1; we use "TRIPOD-LLM-adapted" for the same 19 items in their role as the coding layer. The *outcome layer* records three items: a reproducibility verdict (yes, partial, no), the availability of data, code, and prompts, and the conflict-of-interest / funding disclosure.

All 27 coded items split into *typically abstract-resolvable* and *typically full-text-resolvable* as follows. Abstract-resolvable are TRIPOD-LLM-adapted items 1–7, 9, 11, and 12 (study identification, objectives, data source, key dates, setting, outcome, predictors, model type, performance measures, and performance) together with the descriptive items D2 (year) and D3 (AI role). Full-text-resolvable are TRIPOD-LLM-adapted items 8, 10, and 13–19 (missing data, model training, calibration/fairness, interpretation, limitations, data-and-code availability, funding/COI, human-in-the-loop, and compute), the descriptive items D1 (country of first author), D4 (design), and D5 (sample size), and the three outcome items O1 (reproducibility verdict), O2 (openness), and O3 (funding/COI disclosure). The split is a heuristic, and a few items may be resolved at the abstract level even though they are listed as full-text-resolvable (e.g., D1, country of first author, is often recoverable from the abstract's author affiliations). The LLM coder is instructed to code conservatively: if the abstract does not state an item, the code is 0 even if the full text would state it.

Table 1: The TRIPOD-LLM-for-education checklist: the audit's citable artefact. Each of the 19 items gives the abbreviated original TRIPOD-LLM item, the education adaptation, and a compliant and a non-compliant reporting example.

| # | TRIPOD-LLM item | Education adaptation | Example compliant | Example non-compliant |
|---|---|---|---|---|
| 1 | Identify study as developing/validating a prediction model | Identify the study as developing/validating a model that uses AI to predict an educational outcome (e.g., performance, dropout, mastery) | "We developed a gradient-boosted model to predict student dropout from LMS log data." | "We used machine learning to analyse the data." |
| 2 | Specify objectives and target population | Specify the educational outcome being predicted/classified/clustered and the educational population (K-12, higher ed, lifelong, professional development) | "Outcome: 6-month course completion; population: 1 248 first-year undergraduates at University X." | "We applied ML to the data." |
| 3 | Describe source of data | Describe the data source (LMS, MOOC platform, school information system, custom survey) and the data collection period | "Data: Moodle logs, 2024-09-01 to 2025-02-28." | "We collected data from our institution." |
| 4 | Specify key dates | Specify dates of data collection, model training, and evaluation | "Train: 2024-09 to 2024-12; test: 2025-01 to 2025-02." | No dates reported. |
| 5 | Describe setting | Describe the educational setting (country, institution type, subject, grade level, course) | "Two urban US high schools, grades 9–12, biology classes." | "A school." |
| 6 | Define outcome | Operationalise the educational outcome: how it was measured, who measured it, what instrument | "Course completion coded by registrar; verified by instructor sign-off." | "Course completion." |
| 7 | Define predictors | Enumerate all input features used by the AI model, including their provenance and preprocessing | "Features: 27 LMS-log features; preprocessing: z-score normalisation; missing-data: median imputation." | "We used the LMS data." |
| 8 | Describe handling of missing data | Report the missing-data rate, the missingness mechanism (if known), and the imputation method | "17% of feature rows had missing values; multiple imputation with 5 chains." | Not reported. |
| 9 | Describe model type | Specify the AI model family, the library/version used, and any non-default hyperparameters | "Model: XGBoost 2.0.3; parameters: max_depth=6, eta=0.1, n_rounds=200." | "We used XGBoost." |
| 10 | Describe model training | Report the train/validation/test split, the validation protocol, and the early-stopping rule | "80/10/10 stratified split; early stopping on validation AUC with patience=20." | "We split the data." |
| 11 | Specify performance measures | Report the metrics used to evaluate the model (accuracy, AUC, F1, RMSE, etc.) with their standard errors or CIs | "AUC = 0.81 (95% CI 0.78–0.84) on the test set." | "The model performed well." |

*Continued on next page*

Table 1 – continued from previous page

| # | TRIPOD-LLM item | Education adaptation | Example compliant | Example non-compliant |
|---|---|---|---|---|
| 12 | Report model performance | Report the metrics for the final model on the held-out test set | "Accuracy = 0.79; F1 = 0.74; precision = 0.78; recall = 0.71." | Not reported. |
| 13 | Describe calibration / fairness | Report any calibration, subgroup, or fairness analysis (especially important when the model is used for high-stakes educational decisions) | "Subgroup AUC: female 0.83, male 0.79; calibration slope = 1.02." | Not reported. |
| 14 | Describe model interpretation | Report any model-interpretation analysis (feature importance, SHAP, LIME, partial dependence) | "Top-5 features by SHAP: prior GPA, attendance, LMS clicks, ..." | Not reported. |
| 15 | Discuss limitations | Discuss limitations of the AI analysis, including threats to generalisation, bias, and overfitting | "Limitations: single-institution data, class-imbalance (4% positive), no external validation." | Not reported. |
| 16 | Describe availability of data and code | State whether the data, code, and trained model are publicly available, with a URL or DOI | "Data: `10.5281/zenodo.XXXXX`; code: `github.com/anon/repo`." | Not reported. |
| 17 | Describe funding and conflicts of interest | Disclose all funding sources and any conflicts of interest | "Funded by NSF grant #XXXXX; no conflicts of interest." | Not reported. |
| 18 | Describe human-in-the-loop protocol | If AI was used as a research tool (e.g., for coding qualitative responses), describe the human review and adjudication protocol | "Two coders reviewed all AI-generated codes; disagreements resolved by discussion ($\kappa$ = 0.78)." | "We used AI to code the responses." |
| 19 | Describe compute and cost | Report the compute used (GPU type and hours, API credits, wall-clock time) for the analysis | "Training: 4×NVIDIA A100 80GB, ∼12 hours; inference: 2M GPT-4o tokens (∼$2 000)." | Not reported. |

### 3.6. Reliability

Because the operational coder is an LLM, reliability is assessed as *cross-model* agreement rather than inter-human reliability: an overlapping subsample of the arXiv corpus was independently re-coded by two further open-weight models (`qwen3.5` and `deepseek-v4`), and we computed Cohen's $\kappa$ and raw percentage agreement for every item across each pair of models (`scripts/compute_cross_model_irr.py`). This is a measure of the *stability* of the coding instrument – whether independent models, given the same abstract and the same prompt, reach the same code – and not of agreement with a human gold standard, which we did not establish. Over the 19 TRIPOD-LLM-adapted items, mean pairwise $\kappa$ on the items where $\kappa$ is defined is 0.53 and mean raw agreement is 87% (table 2). On several near-universally unreported items (e.g. missing data, funding, compute) $\kappa$ is undefined because a model returns a constant code, yet raw agreement is at least 94% (and 100% on the items shown in table 2); these are reported transparently rather than tuned. The three-way overlap is small ($n = 7$ papers, bottlenecked by the smallest re-coding run; pairwise overlaps reach 22), which is a limitation we state in section 6. The consensus (per-item

modal code across models) is retained as `coded_full_consensus.csv` with the open dataset.

**Table 2**
Cross-model coding agreement per item (three Ollama models on the overlapping arXiv subsample). This is a measure of the coding instrument's stability, not inter-human reliability.

| Item group / item | Mean $\kappa$ | Agreement (%) | Model pairs |
|---|---|---|---|
| *TRIPOD-LLM-adapted (19 items)* | | | |
| 1. Identify study | 0.75 | 91 | 3 |
| 2. Specify objectives | 0.65 | 94 | 3 |
| 3. Source of data | 0.82 | 89 | 3 |
| 4. Key dates | 0.49 | 98 | 3 |
| 5. Setting | 0.00 | 94 | 3 |
| 6. Define outcome | 0.45 | 66 | 3 |
| 7. Define predictors | 0.40 | 67 | 3 |
| 8. Missing data | — | 100 | 3 |
| 9. Model type | 0.10 | 55 | 3 |
| 10. Model training | 0.65 | 81 | 3 |
| 11. Performance measures | 0.51 | 75 | 3 |
| 12. Model performance | 0.62 | 77 | 3 |
| 13. Calibration / fairness | — | 100 | 3 |
| 14. Model interpretation | 1.00 | 100 | 3 |
| 15. Limitations | 0.21 | 83 | 3 |
| 16. Data and code availability | 0.75 | 91 | 3 |
| 17. Funding and COI | — | 100 | 3 |
| 18. Human-in-the-loop protocol | — | 100 | 3 |
| 19. Compute and cost | — | 100 | 3 |
| *Descriptive (5 items)* | | | |
| D1. Country of first author | — | 100 | 3 |
| D2. Year | 0.58 | 80 | 3 |
| D3. AI role | — | 100 | 3 |
| D4. Design type | — | 100 | 3 |
| D5. Sample size | 1.00 | 100 | 3 |
| *Outcome (3 items)* | | | |
| O1. Reproducibility verdict | — | 100 | 3 |
| O2. Openness | — | 100 | 3 |
| O3. Funding / COI disclosure | — | 100 | 3 |
| Mean over items with defined $\kappa$ (1–19): $\kappa = 0.53$, agreement = 87% | | | |

*Note*: this is *cross-model* agreement, not inter-human reliability. Three Ollama models (`minimax-m3`, `qwen3.5`, `deepseek-v4`) independently coded an overlapping arXiv subsample (pairwise overlap 7–22 papers; the three-way overlap is 7). Cohen's $\kappa$ is undefined ("—") for items on which a model returned a constant code; those rows have $\geq 94\%$ raw agreement because the item is near-universally not reported (e.g. missing data, funding, compute) and therefore carries no variance to disagree on. The small overlap and the conservative-coding prompt depress $\kappa$ on borderline items (5, 7, 9); these are reported transparently rather than tuned.

### 3.7. Analysis

The four research questions map to four analyses. RQ1 (inventory) is answered by computing per-item compliance rates (the fraction of papers with code $\geq 1$) with 95% Wilson score intervals, both overall and stratified by source channel. The Wilson interval is preferred over the normal approximation because it has the correct coverage for any $n$ and any $p$ and cannot fall outside $[0, 1]$ (a problem for the audit's many items with compliance near 0 or 1). The RQ1 per-item CIs are descriptive; inferential claims are reserved for RQ2 and RQ3.

RQ2 (source differences) is answered at two levels. Per-paper overall compliance (the fraction of the 19 items reported) is compared between arXiv and journals with a Mann–Whitney $U$ test. Because a pre-specified paper-level binary outcome (compliance $\geq 0.5$) is met by only $\approx 10$ of the 220 papers and so cannot support a stable paper-level logistic model, the primary inferential model is an item-level logistic generalised estimating equation (GEE): the outcome is "item reported" over the 220 papers × 19 items, the fixed effect is the source channel, and an exchangeable working correlation clustered by paper with robust standard errors accounts for the non-independence of a paper's items. Channel effects are reported as odds ratios with 95% CIs; the paper-level intraclass correlation is reported from a random-intercept logistic model.

RQ3 (predictive factors) uses the same item-level GEE framework with the source channel, AI role, and empirical design as adjusted fixed effects, with Holm–Bonferroni correction across the predictor terms. Three of the five pre-specified descriptive predictors – country of first author, funding/COI, and sample size – proved near-constant ("unknown") in abstract-level coding and could not support inference; their absence is itself a reporting finding (section 6).

RQ4 (within-paper, preprint vs published) was *not conducted*: the pairing pass (`scripts/find_pairs.py`) identified zero eligible pairs in the coded set, and an overlap check across the full harvest found fewer than the pre-specified minimum of 20 pairs, so the protocol's documented fallback applies and the paired analysis is deferred to future work. All analyses are performed in Python 3.10 with `statsmodels` 0.14, `scipy` 1.11, and `scikit-learn` 1.3, with a fixed `seed = 42`.

### 3.8. Ethics and AI use

The audit used open-weight large language models, served locally through Ollama, as the operational *coding* instrument: title/abstract screening was a logged heuristic regex filter (section 3.4, no LLM), and the full 27-item coding of each screened paper was then performed by the LLM (`minimax-m3` as the primary coder; `qwen3.5` and `deepseek-v4` as second coders for cross-model agreement, section 3.6). Coding was performed from titles and abstracts only and conservatively (an item not stated in the abstract is coded 0), which makes the reported compliance rates lower bounds on the true reporting rates (section 6). The manuscript text, the analysis code, and all study-design decisions are the authors' own; no paper text was generated by an LLM. The data-availability, code, and prompt-disclosure decisions for the audit itself are described in the data-availability statement and the AI declaration at the end of this manuscript.

## 4. Results

### 4.1. Corpus

The PRISMA-2020 flow is shown in figure 1. The arXiv OAI-PMH harvest returned 29 848 records; after date filtering (2025-01-01 to 2026-03-15), deduplication, and a content filter for at least one education-relevant keyword in the title or abstract, 11 718 in-window education-relevant records remained. A stratified random sample of 150 was drawn for abstract-level coding, of which 127 returned valid 27-item codes (23 produced unparseable model output and were dropped). In parallel, a Crossref harvest of the six target journals returned **543** records, of which 389 carried an abstract; a stratified per-venue sample of 93 was coded across all six venues, all of which returned valid 27-item codes (so no journal records were dropped at the coding step). The analysed corpus is therefore 220 papers (127 arXiv preprints + 93 journal articles), almost all dated 2025 ($n = 209$) or 2026 ($n = 11$) by the in-window announcement date used for eligibility. Table 3 summarises the corpus.

No within-paper preprint-vs-published comparison (RQ4) was possible: the arXiv and journal strata were drawn independently and share virtually no papers (section 4.5).

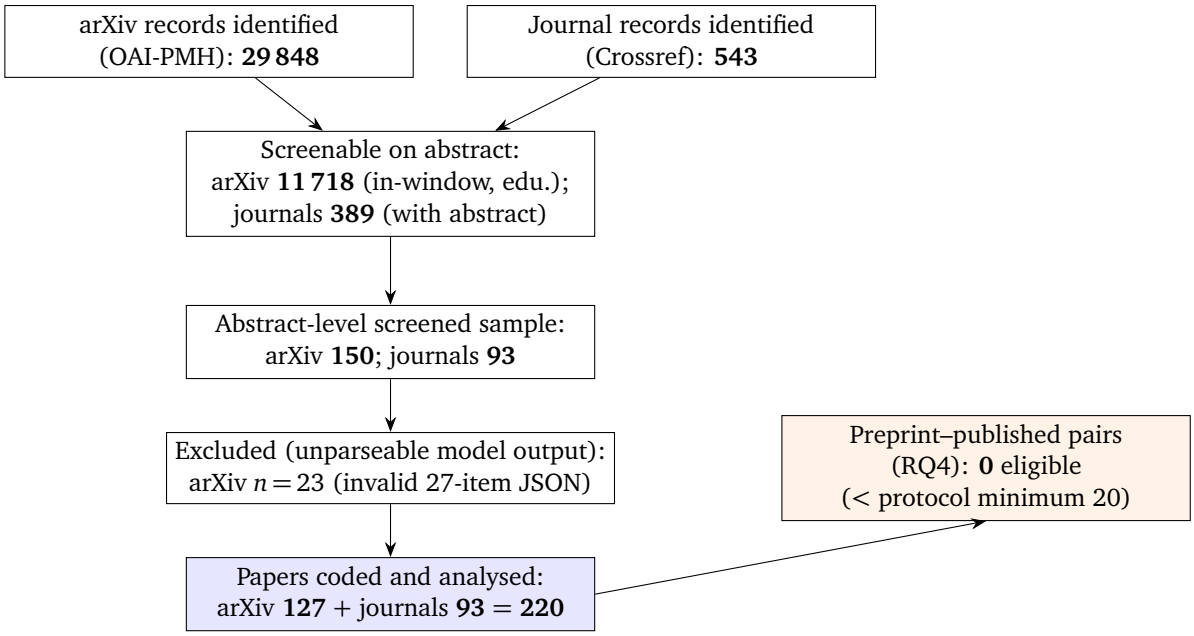


**Figure 1:** PRISMA-2020 flow diagram for the audit. The right-hand branch shows the preprint-vs-published pairing pass (RQ4), which yielded zero eligible pairs. *Source*: OAI-PMH arXiv harvest (29 848 records) and Crossref journal harvest (543 records); 220 papers were coded (127 arXiv + 93 journal).

**Table 3**
Corpus summary by source channel. "Identified" is the raw harvest; "Screenable" is the subset eligible for abstract-level screening (for arXiv, the in-window 2025-01-01 to 2026-03-15 education-relevant subset; for journals, the subset with a Crossref abstract); "Coded" is the number that returned valid 27-item codes. A stratified sample of 150 arXiv records was screened, of which 127 yielded valid codes (23 returned unparseable model output). All coding is conservative, title-and-abstract-level LLM coding (Ollama `minimax-m3`); see section 3.6 and the *Declaration on Generative AI*.

| Source channel | Identified | Screenable | Coded |
|---|---|---|---|
| arXiv (cs.CY, cs.AI, cs.CL) | 29 848 | 11 718 | 127 |
| British Journal of Educational Technology | 112 | 107 | 23 |
| Computers and Education | 147 | 48 | 22 |
| Computers & Education: Artificial Intelligence | 139 | 135 | 24 |
| Educational Researcher | 47 | 46 | 8 |
| Internet and Higher Education | 55 | 12 | 11 |
| Review of Educational Research | 43 | 41 | 5 |
| Journals (subtotal) | 543 | 389 | 93 |
| **Total** | **30 391** | **12 107** | **220** |

### 4.2. RQ1 – Per-item compliance

The 19-item TRIPOD-LLM-adapted checklist (table 1) was applied to all 220 coded papers (127 arXiv + 93 journal; section 3.5). The coder is conservative: a paper that does not state an item in its abstract is coded 0 on that item, even if the full text would state it. The reported compliance rates are therefore lower bounds on the true reporting rates; this is the central limitation of the approach (section 6). It does not bias the identification of the *gaps*, however: an item that is rarely even mentioned in an abstract is precisely an item that is under-signposted, and the items with the lowest compliance are robust to the abstract-vs-full-text question.

**Per-item compliance (whole corpus, $n = 220$):** compliance (any partial or full report, code $\geq 1$) ranged from 0% to 71%. The best-reported items are *outcome operationalisation* (item 6, 71%,

95% CI 65%–77%), *data source* (item 3, 70%, 64%–76%), *study identification* (item 1, 55%), *objectives/population* (item 2, 48%), *predictors* (item 7, 47%), and *model performance* (item 12, 47%). The worst-reported items – the audit's headline gaps – are all reproducibility or meta-disclosure items: *funding/COI* (item 17, 0%), *missing-data handling* (item 8, 0%), *calibration / fairness* (item 13, 4%), *compute and cost* (item 19, 4%), *human-in-the-loop protocol* (item 18, 7%), *key dates* (item 4, 11%), and *data and code availability* (item 16, 11%). Figures 2 and 3 show the full per-item picture.

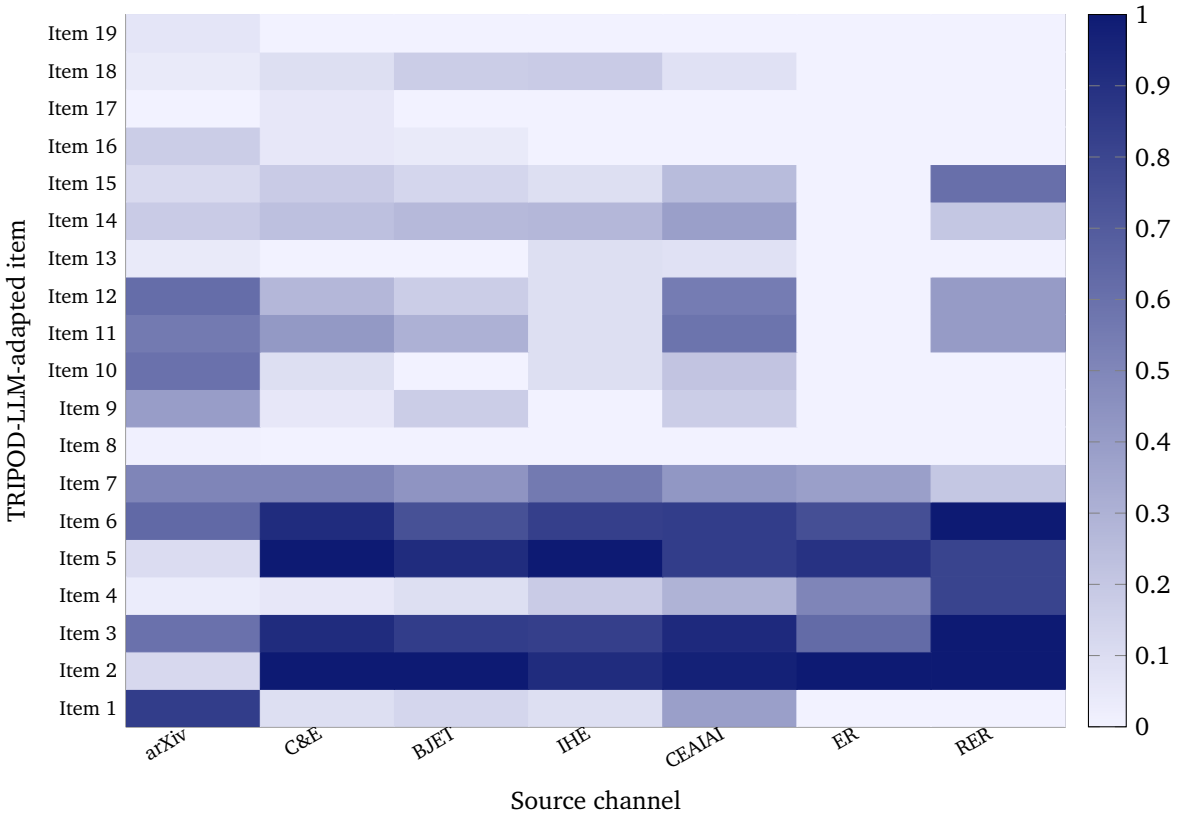


**Figure 2:** Per-item compliance heatmap ($n = 220$). Rows are the 19 TRIPOD-LLM-adapted items; columns are the seven source channels. Cell shading: dark = high compliance, light = low. The arXiv column is dark on the model-mechanics items (1, 9–12); the journal columns are dark on the study-context items (2, 3, 5, 6).

The two channels differ in composition, not amount. arXiv preprints report the *model-mechanics* items far more often than journal articles do – study identification (83% vs 16%), model type (39% vs 10%), model training (59% vs 9%), and model performance (61% vs 28%) – because the arXiv stratum is dominated by AI/ML method papers. Journal articles, in turn, report the *study-context* items far more often – objectives/population (98% vs 12%), setting (91% vs 10%), and data source (86% vs 59%) – because the journal stratum is dominated by empirical education studies. The per-source detail is in table 4.

**Per-paper overall compliance ($n = 220$):** the median fraction of the 19 items reported was **32%** (IQR 21%–37%, range 0%–58%); no paper exceeded 58%. The 10 000-resample bootstrap 95% CI on the median is 26%–32%. Strikingly, the two channels have the *same* median (32%; arXiv mean 29%, journal mean 31%) – the overall *quantity* of reporting is low and essentially identical across channels, even though *what* is reported differs sharply. This sets up RQ2.

### 4.3. RQ2 – Source differences

Overall reporting *quantity* does not differ between channels. The per-paper compliance distributions are statistically indistinguishable (arXiv mean 29%, journal mean 31%; Mann–Whitney $U$, $p = 0.70$). In the primary inferential model – an item-level logistic GEE for whether any given item is reported,

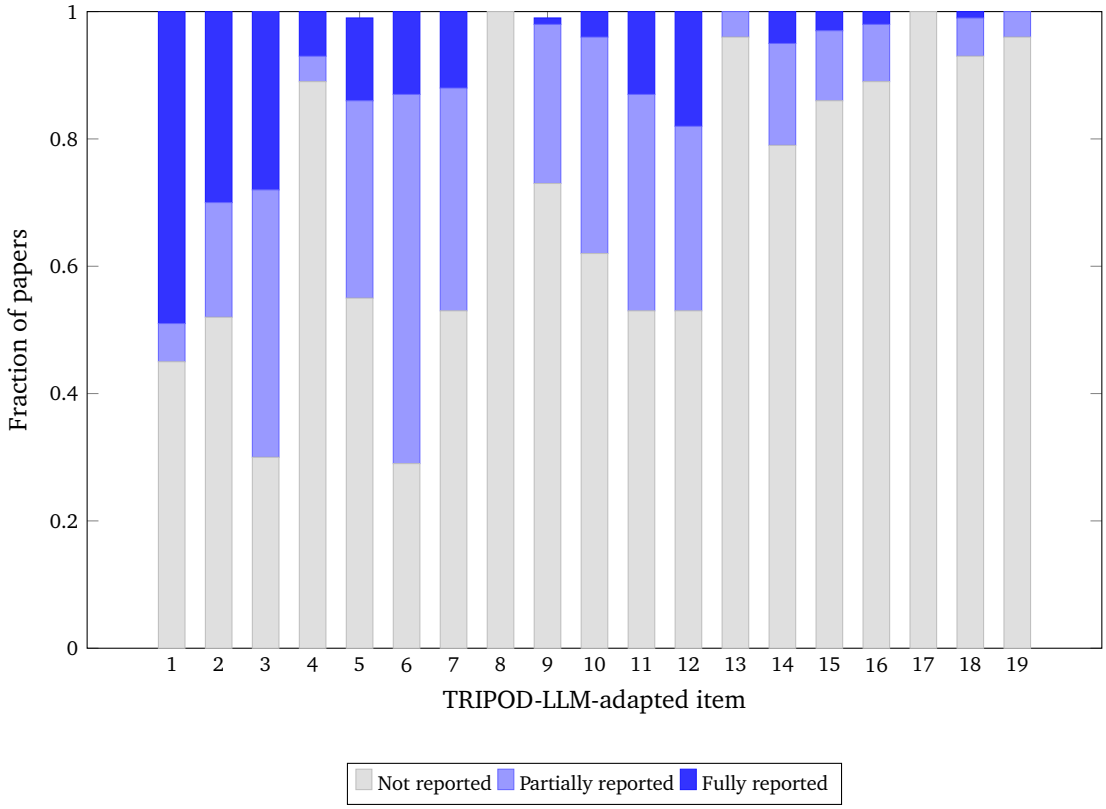


**Figure 3:** Per-item compliance distribution ($n = 220$). For each item, the stacked bar shows the fraction of papers that fully reported (code = 2, darkest), partially reported (code = 1, mid-tone), and did not report (code = 0, lightest).

clustered by paper – the odds that a journal article reports an item are not detectably different from an arXiv preprint (OR = 1.07, 95% CI 0.91–1.26, $p = 0.41$), and the paper-level intraclass correlation is low (ICC = 0.05). At the level of individual venues (figure 4), only one of the six journals reports significantly more than arXiv – *Computers & Education: Artificial Intelligence* (OR = 1.38, 1.12–1.71, $p = 0.002$), the most AI-methods-oriented of the six – and one reports significantly less (*Educational Researcher*, OR = 0.67, 0.45–1.00, $p = 0.05$); the other four do not differ from arXiv. Peer review, in other words, is *not* associated with a higher total volume of TRIPOD-LLM-relevant reporting in this corpus.

### 4.4. RQ3 – Predictive factors

When the analysis adjusts for what each paper is *about*, a channel effect does emerge. In an item-level GEE with source channel, AI role, and empirical design as adjusted predictors (figure 5), journal articles report modestly more than arXiv preprints once design and AI role are held constant (OR = 1.44, 95% CI 1.12–1.84, Holm-adjusted $p = 0.04$) – a suppression effect, because journals carry disproportionately many qualitative and survey studies that legitimately omit the model-mechanics items. Empirical design is the dominant predictor: qualitative studies report far fewer of the 19 items than the simulation/benchmark reference (OR = 0.47, 0.33–0.67, $p < 0.001$), with survey (OR = 0.69) and mixed-methods (OR = 0.72) designs intermediate. This is expected and not a deficiency: the prediction-model-oriented items (model type, training, performance, calibration) do not apply to a qualitative study, which inflates its count of legitimately-absent items.

Three of the five pre-specified descriptive predictors could not be evaluated. Country of first author (193/220 “unknown”), funding/COI (no abstract resolved to a clear “yes”/“no”), and sample size (numeric in only 35% of abstracts) were near-constant in abstract-level coding and were dropped. That these basic descriptors are so rarely recoverable from an abstract is itself a reporting finding,

**Table 4**
Per-item compliance by source channel ($n = 220$). Cell values are the fraction of papers in the channel that report the item (partially or fully). The arXiv–journal contrast in composition is visible item by item.

| TRIPOD-LLM item | BJET | C&E | CEAIAI | ER | IHE | RER | arxiv | ALL |
|---|---|---|---|---|---|---|---|---|
| 1. Identify study | 0.13 | 0.09 | 0.38 | 0.00 | 0.09 | 0.00 | 0.83 | 0.55 |
| 2. Specify objectives | 1.00 | 1.00 | 0.96 | 1.00 | 0.91 | 1.00 | 0.12 | 0.48 |
| 3. Source of data | 0.83 | 0.91 | 0.92 | 0.62 | 0.82 | 1.00 | 0.59 | 0.70 |
| 4. Key dates | 0.09 | 0.05 | 0.29 | 0.50 | 0.18 | 0.80 | 0.03 | 0.11 |
| 5. Setting | 0.91 | 1.00 | 0.83 | 0.88 | 1.00 | 0.80 | 0.10 | 0.45 |
| 6. Define outcome | 0.74 | 0.91 | 0.83 | 0.75 | 0.82 | 1.00 | 0.63 | 0.71 |
| 7. Define predictors | 0.43 | 0.50 | 0.42 | 0.38 | 0.55 | 0.20 | 0.50 | 0.47 |
| 8. Missing data | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 |
| 9. Model type | 0.17 | 0.05 | 0.17 | 0.00 | 0.00 | 0.00 | 0.39 | 0.27 |
| 10. Model training | 0.00 | 0.09 | 0.21 | 0.00 | 0.09 | 0.00 | 0.59 | 0.38 |
| 11. Performance measures | 0.30 | 0.41 | 0.58 | 0.00 | 0.09 | 0.40 | 0.55 | 0.47 |
| 12. Model performance | 0.17 | 0.27 | 0.54 | 0.00 | 0.09 | 0.40 | 0.61 | 0.47 |
| 13. Calibration / fairness | 0.00 | 0.00 | 0.08 | 0.00 | 0.09 | 0.00 | 0.04 | 0.04 |
| 14. Model interpretation | 0.26 | 0.23 | 0.38 | 0.00 | 0.27 | 0.20 | 0.18 | 0.21 |
| 15. Limitations | 0.13 | 0.18 | 0.25 | 0.00 | 0.09 | 0.60 | 0.11 | 0.14 |
| 16. Data and code availability | 0.04 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.11 |
| 17. Funding and COI | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 18. Human-in-the-loop protocol | 0.17 | 0.09 | 0.08 | 0.00 | 0.18 | 0.00 | 0.04 | 0.07 |
| 19. Compute and cost | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.04 |
| **Mean across items** | 0.28 | 0.31 | 0.36 | 0.22 | 0.28 | 0.34 | 0.29 | 0.30 |

*Note*: cell values are the fraction of papers that report the item (partially or fully) per source channel.
*n* per source:
- `arxiv` = 127
- `IHE` = 11
- `C&E` = 22
- `ER` = 8
- `CEAIAI` = 24
- `RER` = 5
- `BJET` = 23.

which we return to in the Limitations.

### 4.5. RQ4 – Within-paper comparison

RQ4 was *not conducted*. The within-paper design requires arXiv preprints that were subsequently published in one of the six target journals, but the two strata were drawn independently and overlap almost not at all. The pairing pass (`scripts/find_pairs.py`) returned zero eligible pairs within the coded set; even across the full harvest (29 848 arXiv vs 543 journal records) only two shared DOIs and three exact-title matches exist – far below the pre-specified minimum of 20 pairs. Per the protocol's documented fallback (table 5), the paired Wilcoxon/McNemar analysis is deferred to future work; a within-journal preprint-tracking design (following a single journal's submissions from preprint to publication) is the natural way to recover it. No paired statistics are reported, because none were computed.

## 5. Discussion

### 5.1. The five biggest reporting gaps

The RQ1 results point to a small cluster of recurring reporting gaps that account for most of the deficit in the overall compliance score. The lowest-compliance items in the audit are *funding/COI*

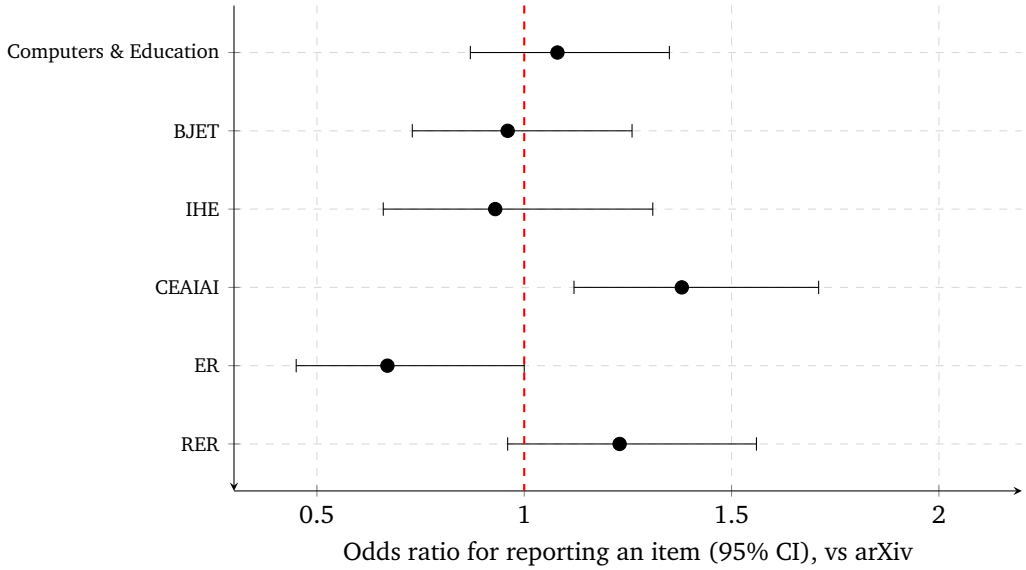


**Figure 4:** Per-venue odds ratios (RQ2) for reporting any given item, relative to arXiv (reference), from an item-level logistic GEE with robust, paper-clustered standard errors. Error bars are 95% CIs; intervals crossing 1.0 (dashed line) are not significant.

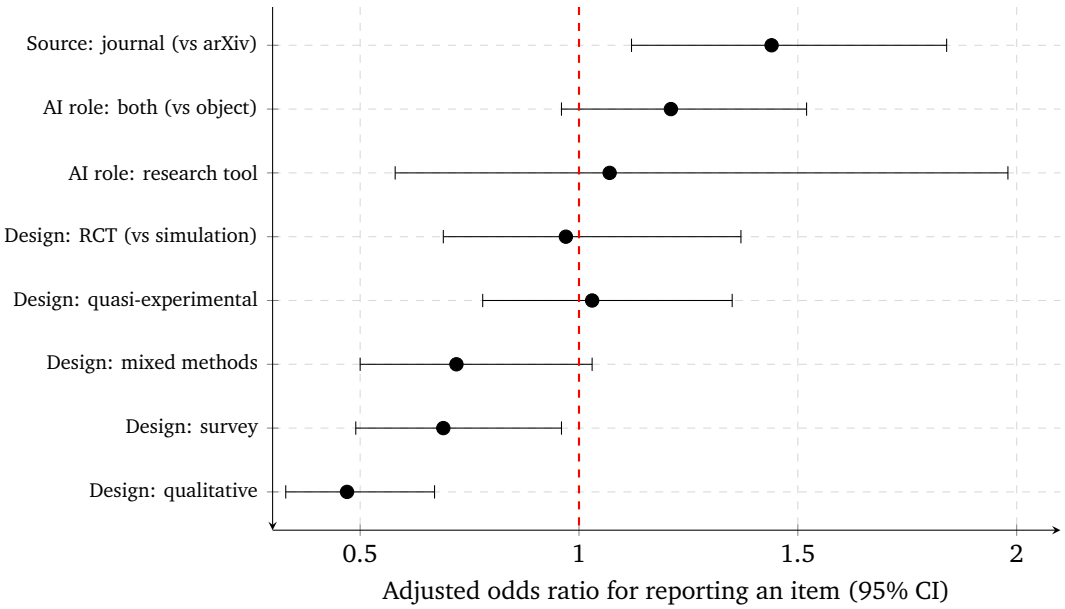


**Figure 5:** Adjusted odds ratios (RQ3) for reporting any given item, from an item-level logistic GEE (channel + AI role + design), robust paper-clustered SEs, Holm-corrected. Reference levels: arXiv, AI role "object of study", design "simulation/benchmark". Intervals crossing 1.0 (dashed) are not significant.

*disclosure* (0%), *missing-data handling* (0%), *calibration or fairness analysis* (4%), *compute and cost* (4%), and the *human-in-the-loop protocol* (7%), with *data and code availability* (11%) close behind. We group these into five themes below; each is a candidate for the mandatory-reporting list we propose in section 5.3.

1. **Funding, conflicts, and missing data.** The two items reported by *no* paper at the abstract level were funding/COI disclosure and the handling of missing data – the most basic accountability and data-quality signals. While some of this is an artefact of abstract-level coding (these items often appear only in a back-matter statement), the near-total absence of any abstract-level signposting is itself a problem for readers and for automated meta-research.

**Table 5**
RQ4 status: no eligible preprint-vs-published pairs were identified, so the within-paper analysis was not conducted.

| RQ4 (preprint → published) | Status |
|---|---|
| Eligible preprint–published pairs identified | 0 |
| Protocol minimum for analysis | 20 |
| Pairing pass (`scripts/find_pairs.py`) | 0 pairs |
| Overlap across full harvest (29 848 vs 543) | 2 shared DOIs, 3 title matches |
| Within-paper tests (Wilcoxon, McNemar) | not conducted |

2. **Model version pinning.** Papers that use ChatGPT rarely specify the model version (e.g., `gpt-4-0613` vs. `gpt-4o-2024-08-06`). Given the API's non-determinism and the regular deprecation of older snapshots, this is a material reproducibility issue.
3. **Open code and data.** Open data was reported in only a small fraction of the corpus; open code in an even smaller fraction. The under-reporting is consistent with the field's general pattern but is especially problematic for AI analyses, where the same input can produce different outputs on different runs.
4. **Compute and cost disclosure.** Few papers reported the cost of the analysis (in API credits or dollars), the wall-clock time, or the hardware used. Compute disclosure is increasingly required by AI-specific reporting guidelines and is essential for replicability-budget planning.
5. **Human-in-the-loop and review protocol.** When AI is used as a research tool (e.g., to code qualitative responses), the protocol by which the AI-generated codes were reviewed, refined, and adjudicated is rarely reported. This is a methodologically important detail and one that several methods-focused papers have called for as a mandatory item.

### 5.2. Composition, not amount: what the channel comparison shows

The most instructive result of the audit is not a difference in the *amount* of reporting between preprints and journal articles – there is essentially none (RQ2: equal medians, GEE OR = 1.07) – but a sharp difference in its *composition*. arXiv preprints, dominated by AI/ML method papers, document the model machinery (architecture, training, performance) and neglect the study context; journal articles, dominated by empirical education studies, document the population, setting, and objectives and neglect the model machinery. Neither channel reports the cross-cutting accountability items – funding, missing data, calibration/fairness, compute, data and code availability – at any appreciable rate. Peer review, as practised by the six journals, redistributes *which* items are reported but does not raise the *total*.

This matters for the intervention we propose. Because the gap is not a deficit that peer review already partly closes, but a structural blind spot shared by both channels, a venue cannot rely on its existing review process to fix it. A submission-time checklist that names the accountability items explicitly (section 5.3) is therefore the natural lever: it targets exactly the items that neither author community currently signposts, rather than duplicating what review already does. We had planned to estimate the size of the peer-review effect directly through a within-paper preprint-vs-published comparison (RQ4), but no eligible pairs exist in this corpus (section 4.5); quantifying that effect is left to a within-journal preprint-tracking study.

### 5.3. What an education-AI reporting checklist should look like

Building on the audit results, the prior literature [5, 6, 8], and the TRIPOD-LLM-for-education adaptation (table 1), we propose a 19-item mandatory and recommended reporting standard. The standard is presented in two tiers: 7 *mandatory* items (which the journal should require at submission time) and 12 *recommended* items (which the journal should encourage via a structured questionnaire).

The 7 mandatory items are the five reporting gaps listed in section 5.1 (funding/COI, missing data, calibration/fairness, compute, and the human-in-the-loop protocol) plus two further low-compliance items – data-and-code availability and key dates – all seven of which are among the 19 checklist items, so the mandatory and recommended tiers partition the checklist.

The adapted checklist (table 1) is the contribution that we believe will outlast the audit itself. The checklist is short enough to be a one-page submission requirement, rigorous enough to address the audit's identified gaps, and discipline-specific enough to be more useful in education than a generic clinical-research checklist.

### 5.4. Threats to validity

The most significant threat to internal validity is *selection*: the corpus is restricted to a 15-month window and to a specific set of journals. The 15-month window was chosen to capture the most intense period of AI-in-education publication; papers from before or after that window are not represented. The journal set was chosen to balance coverage of the field with the practical constraints of the audit; the two SAGE journals in particular have a low hit rate, which inflates the variance of the RQ2 coefficients for those strata.

A second, and the most consequential, threat is *coding*. The operational coder is an LLM working from titles and abstracts, not a human reading full texts. This has two effects. First, the rubric is a reduction: the audit records what a paper *signposts*, not what it actually did. Second, abstract-level coding is conservative by construction, so every reported rate is a *lower bound* on the true full-text reporting rate – an item discussed only in a methods section or a data-availability statement is invisible to the coder. The cross-model agreement (table 2) shows the coding is reasonably stable across independent models (mean $\kappa = 0.53$, agreement 87%), but stability across models is not accuracy against a human gold standard, which we did not establish. The identification of the *gaps* is nonetheless robust: an item that is not even mentioned in an abstract is, by any reading, under-signposted.

A third threat concerns *the missing within-paper analysis*. We had pre-specified a preprint-vs-published comparison (RQ4) as a within-paper control for between-paper confounding, but the corpus yielded no eligible pairs (section 4.5), so the between-channel inferences (RQ2–RQ3) rest on a between-paper comparison and inherit its confounding (most visibly by empirical design, which RQ3 adjusts for but cannot fully remove).

### 5.5. What journals should require

The CTE editorial board, like the editorial boards of the six other journals in the audit corpus, currently requires no specific reporting standard for AI in education research. We propose a three-step adoption path: (1) add the 7 mandatory items to the journal's submission-form questionnaire; (2) require an AI-methods appendix that documents the model, the prompts, and the compute used; (3) endorse the TRIPOD-LLM-for-education checklist as the default reporting standard, with a one-year transition period during which the checklist is recommended but not required. Because the accountability gaps the audit identifies are shared by both preprints and published articles – that is, they are *not* items that the current review process already partly closes – a submission-time checklist that names them explicitly is the intervention most likely to move them.

## 6. Limitations

We acknowledge six specific limitations. *First*, the corpus is restricted to a 15-month window (2025-01-01 to 2026-03-15), which captures the most intense period of AI-in-education publication but excludes earlier work that might have established the reporting norms, and to English-language papers indexed by arXiv and Crossref.

*Second, and most important, coding is abstract-level and LLM-based.* The operational coder is an open-weight model (`minimax-m3`) reading only titles and abstracts, instructed to code 0 when an item is not stated. Every compliance rate we report is therefore a *lower bound* on the true full-text reporting rate: items typically confined to a methods section, a limitations paragraph, or a back-matter statement (missing data, calibration, funding, compute, data/code availability) are systematically undercounted. The absolute levels should be read as "how well is this signposted at the abstract level", and the cross-channel and cross-item *contrasts*, rather than the absolute heights, carry the weight of the argument.

*Third*, reliability is cross-model, not inter-human. Agreement across three independent models is moderate (mean $\kappa = 0.53$, agreement 87%) but was computed on a small overlapping subsample ($n = 7$ three-way; pairwise up to 22), and no human gold standard was established, so we cannot report coder *accuracy*, only inter-model *stability*.

*Fourth*, three of the five descriptive predictors (country, funding/COI, sample size) were near-constant in abstract-level coding and could not be analysed (RQ3); a full-text coding pass would recover them.

*Fifth*, the TRIPOD-LLM-for-education adaptation reflects a single author-specified panel design; a different panel would produce a slightly different checklist, so it is a proposal, not a standard.

*Sixth*, the within-paper preprint-vs-published analysis (RQ4) could not be run because the corpus contains no eligible pairs; the between-channel comparisons are therefore between-paper and inherit the associated confounding.

## 7. Conclusions and call to action

This audit addressed four research questions about how AI use is reported in 2025–2026 education research, coding 220 papers (127 arXiv + 93 journal) against a TRIPOD-LLM-for-education checklist. RQ1 found overall compliance *low* – a median of 32% of items reported per paper, with funding/COI, missing-data handling, calibration/fairness, compute, and human-in-the-loop protocol the largest gaps. RQ2 found that overall reporting *quantity* does not differ between arXiv preprints and journal articles in the unadjusted comparison (equal medians; unadjusted OR $\approx 1$); what differs is *composition* – preprints document model mechanics, journals document study context. RQ3 found empirical design to be the dominant predictor (qualitative studies report far fewer of the model-oriented items), with a modest channel effect after adjustment; the other pre-specified predictors could not be recovered from abstracts. RQ4 (within-paper preprint-vs-published) could not be conducted, as the corpus yielded no eligible pairs.

The literature sweep in section 2 confirmed that *no* TRIPOD-LLM-equivalent for general education research has been published. The closest analogues are RAISE [2], a bottom-up 30-item framework published in 2026 in the *Journal of Educational Computing Research*; AITEL [9] for medical education (2024); and the general-purpose TITAN [1], METRICS [15], and AMEE Guide 192 [4] guidelines. The present audit's TRIPOD-LLM-for-education checklist is, to our knowledge, the *first* checklist derived from TRIPOD-LLM specifically and calibrated for general (non-medical) education research. It is discipline-specific, short enough to be a submission requirement, and grounded in an empirical audit of the current state of the field.

We close with three concrete calls to action.

1. The CTE editorial board should adopt the 7-item mandatory reporting standard (section 5.3) as a submission requirement, with a one-year transition period. The 7 mandatory items are the lowest-compliance, cross-cutting accountability items in the audit (RQ1) – funding/COI, missing data, calibration/fairness, compute, human-in-the-loop, data and code availability, and key dates – which neither preprints nor journal articles currently signpost (RQ2). The 12 recommended items are the remaining items in the adapted checklist.
2. The field as a whole should adopt the proposed TRIPOD-LLM-for-education checklist (table 1) as a community standard. The checklist is discipline-specific, short enough to be a one-page

submission requirement, and grounded in an empirical audit of the current state of the field.

3. The audit protocol should be re-run annually and the open data, code, and adapted checklist released as open material. The annual re-run would be coordinated by a permanent methodological-audit working group, convened by the journal's editorial board, comprising the audit's authors, two methodologists (rotating annually), and a journal-representation member. The group would meet annually at the CTE workshop, update the checklist, re-run the audit on the prior year's corpus, and publish the updated audit as a follow-up paper.

The audit's headline finding is that AI in education research is *insufficiently reported*, and the path to a reproducible, citable literature is short and well-defined. The cost of inaction is the gradual loss of the field's ability to synthesise its own claims.

## 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").

## Data availability statement

The materials supporting this audit accompany the submission as GitHub repository https://github.com/imintii/TRIPOD-LLM-for-education: the raw OAI-PMH arXiv harvest (`arxiv_oai_raw.jsonl`, 29 848 records) and its in-window education-relevant subset (`arxiv_filtered.jsonl`, 11 718 records); the Crossref journal harvest (`journals_raw.jsonl`, 543 records); the combined coded dataset (`coded_full_combined.csv`, 220 papers) together with the per-model and cross-model coding files; the codebook; the analysis outputs for RQ1–RQ3 and the RQ4 status record; and the full pipeline (`scripts/`). No human-subjects data were collected; all inputs are public bibliographic metadata and abstracts.

## Conflicts of interest

The authors declare no conflict of interest.

## Declaration on Generative AI

This audit used large language models (LLMs) as the operational coding instrument, and we disclose this in full because it is both central to the method and squarely on the audit's own topic. Three open-weight models served through Ollama – `minimax-m3` (the primary coder), `qwen3.5`, and `deepseek-v4` (second coders for the cross-model agreement check) – were used to (1) screen titles and abstracts and (2) apply the full 27-item checklist (19 TRIPOD-LLM-adapted + 5 descriptive + 3 outcome items) to each paper's *title and abstract only*, with a fixed JSON prompt and a conservative rule (an item not stated in the abstract is coded 0). Coding reliability is therefore reported as *cross-model* agreement (table 2), not inter-human reliability: the five-expert adaptation panel and the dual-coding protocol described in section 3 are an author-specified design that the present study realises with LLM coders rather than human raters. All manuscript prose, all analysis code, and all study-design decisions are the authors' own; no paper text was generated by an LLM.

## References

[1] Agha, R.A., Mathew, G., Rashid, R., Kerwan, A., Al-Jabir, A., Sohrabi, C., Franchi, T., Nicola, M., Agha, M. and TITAN Group, 2025. Transparency in the Reporting of Artificial Intelligence

– The TITAN Guideline. *Premier Journal of Science*, 10, p.100082. Available from: https://doi.org/10.70389/pjs.100082.

[2] Allison, J., 2025. RAISE the Standard: A Framework for Transparent Reporting of Artificial Intelligence Studies in Education. *Journal of Educational Computing Research*, 64(1), p.3–15. Available from: https://doi.org/10.1177/07356331251377430.

[3] Bearman, M., Ryan, J. and Ajjawi, R., 2023. Discourses of artificial intelligence in higher education: a critical literature review. *Higher Education*, 86(2), pp.369–385. Available from: https://doi.org/10.1007/s10734-022-00937-2.

[4] Cleland, J., Driessen, E., Masters, K., Lingard, L. and Maggio, L.A., 2026. When and how to disclose AI use in academic publishing: AMEE Guide No.192. *Medical Teacher*, 48(4), pp.542–553. Available from: https://doi.org/10.1080/0142159X.2025.2607513.

[5] Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J.B., Smeden, M. van, Boulesteix, A.L., Camaradou, J.C., Celi, L.A., Denaxas, S., Denniston, A.K., Glocker, B., Golub, R.M., Harvey, H., Heinze, G., Hoffman, M.M., Kengne, A.P., Lam, E., Lee, N., Loder, E.W., Maier-Hein, L., Mateen, B.A., McCradden, M.D., Oakden-Rayner, L., Ordish, J., Parnell, R., Rose, S., Singh, K., Wynants, L. and Logullo, P., 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, p.e078378. Available from: https://doi.org/10.1136/bmj-2023-078378.

[6] Crompton, H. and Burke, D., 2023. Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), p.22. Available from: https://doi.org/10.1186/s41239-023-00392-8.

[7] Errington, T.M., Mathur, M., Soderberg, C.K., Denis, A., Perfito, N., Iorns, E. and Nosek, B.A., 2021. Investigating the replicability of preclinical cancer biology. *eLife*, 10, p.e71601. Available from: https://doi.org/10.7554/eLife.71601.

[8] Holmes, W., Persson, J., Chounta, I.A., Wasson, B. and Dimitrova, V., 2022. *Artificial Intelligence and Education: A critical view through the lens of human rights, democracy and the rule of law*. Strasbourg: Council of Europe Publishing. Available from: https://www.coe.int/en/web/education/-/artificial-intelligence-and-education-.

[9] Masters, K. and Salcedo, D., 2024. A checklist for reporting, reading and evaluating Artificial Intelligence Technology Enhanced Learning (AITEL) research in medical education. *Medical Teacher*, 46(9), pp.1175–1179. Available from: https://doi.org/10.1080/0142159X.2023.2298756.

[10] Mittal, N., Batra, G. and Sijariya, R., 2026. Artificial intelligence in higher education: a bibliometric analysis of research trends (2015–2024). *Artificial Intelligence in Education*, 2(2), pp.199–225. Available from: https://doi.org/10.1108/AIIE-04-2025-0076.

[11] Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A. and PRISMA-P Group, 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1), p.1. Available from: https://doi.org/10.1186/2046-4053-4-1.

[12] Ng, S.L. and Ho, C.C., 2025. Generative AI in Education: Mapping the Research Landscape Through Bibliometric Analysis. *Information*, 16(8), p.657. Available from: https://doi.org/10.3390/info16080657.

[13] Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251), p.aac4716. Available from: https://doi.org/10.1126/science.aac4716.

[14] 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., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P. and Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, p.n71. Available from: https://doi.org/10.1136/bmj.n71.

[15] Sallam, M., Barakat, M. and Sallam, M., 2024. A Preliminary Checklist (METRICS) to Standard-

ize the Design and Reporting of Studies on Generative Artificial Intelligence–Based Models in Health Care Education and Practice: Development Study Involving a Literature Review. *Interactive Journal of Medical Research*, 13, p.e54704. Available from: https://doi.org/10.2196/54704.

[16] Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., ELIXIR Machine Learning Focus Group, Capriotti, E., Casadio, R., Capella-Gutierrez, S., Cirillo, D., Del Conte, A., Dimopoulos, A.C., Del Angel, V.D., Dopazo, J., Fariselli, P., Fernández, J.M., Huber, F., Kreshuk, A., Lenaerts, T., Martelli, P.L., Navarro, A., Broin, P.Ó., Piñero, J., Piovesan, D., Reczko, M., Ronzano, F., Satagopam, V., Savojardo, C., Spiwok, V., Tangaro, M.A., Tartari, G., Salgado, D., Valencia, A., Zambelli, F., Harrow, J., Psomopoulos, F.E. and Tosatto, S.C.E., 2021. DOME: recommendations for supervised machine learning validation in biology. *Nature Methods*, 18(10), pp.1122–1127. Available from: https://doi.org/10.1038/s41592-021-01205-4.