Supplement S1: Search Provenance, Complete Retrieval Ledgers, Forensic Numerical Audit, Effect-Size Derivations, and Result-Specific Critical Appraisals

Author: Alexander Kuznetsov, Ph.D. (5 Steps Academy, Singapore; ORCID: 0009-0002-6652-3544)
Companion Manuscript: *Independent Learning after Generative AI Support in School Mathematics: A Systematic Review and Assessment-Stratified Meta-Analysis*

S1.1 Two-Stream Search Provenance, Exact Boolean Queries, and PRISMA 2020 Reconciliation

All rights reserved. © 2026 Alexander Kuznetsov. Personal reading, downloading, and scholarly citation are permitted; redistribution, commercial reuse, or adaptation requires prior written permission from the copyright holder unless permitted by law.
No prospective registration is claimed. Candidate discovery united two documented bibliographic streams spanning 1 January 2022 through 8 November 2026: our core database register (Stream 1) and programmatic multi-index expansion paired with bidirectional citation chasing (Stream 2).

S1.1.1 Stream 1: Core ERIC (E01–E08), APA PsycNet, and Scoping Seed Register

Eight structured ERIC component queries ran on 8 October 2026 and returned 451 raw occurrences across 299 unique ERIC identifiers. Every query paired a generative model term with mathematics curriculum terms and publication years 2022–2026:

- **E01**: ChatGPT AND (math* OR algebra OR geometr* OR arithmetic OR calculus OR "word problem" OR "mathematical reasoning") AND publicationdateyear:(2022 TO 2026)
- **E02**: "generative AI" AND (math* OR algebra OR geometr* OR arithmetic OR calculus OR "word problem" OR "mathematical reasoning") AND publicationdateyear:(2022 TO 2026)
- **E03**: "generative artificial intelligence" AND (math* OR algebra OR geometr* OR arithmetic OR calculus OR "word problem" OR "mathematical reasoning") AND publicationdateyear:(2022 TO 2026)
- **E04**: "large language model" AND (math* OR algebra OR geometr* OR arithmetic OR calculus OR "word problem" OR "mathematical reasoning") AND publicationdateyear:(2022 TO 2026)
- **E05**: ("GPT-3" OR "GPT-4" OR "GPT-4o") AND (math* OR algebra OR geometr* OR arithmetic OR calculus OR "word problem" OR "mathematical reasoning") AND publicationdateyear:(2022 TO 2026)
- **E06**: (Claude OR Gemini OR Llama OR Copilot) AND (math* OR algebra OR geometr* OR arithmetic OR calculus OR "word problem" OR "mathematical reasoning") AND publicationdateyear:(2022 TO 2026)
- **E07**: ("GPT-3.5" OR "GPT-5" OR "generative pretrained transformer") AND (math* OR algebra OR geometr* OR arithmetic OR calculus OR "word problem" OR "mathematical reasoning") AND publicationdateyear:(2022 TO 2026)
- **E08**: (LLM OR LLMs OR "large language models" OR Bard) AND (math* OR algebra OR geometr* OR arithmetic OR calculus OR "word problem" OR "mathematical reasoning") AND publicationdateyear:(2022 TO 2026)

APA PsycNet was queried on 14 October 2026 across Any Field for publication years 2022–2026 (759 raw hits; 758 unique records). One Boolean conjunction joined two concept blocks:

- **Block 1 (GenAI):** (ChatGPT OR "Chat GPT" OR "generative AI" OR "generative artificial intelligence" OR GenAI OR "large language model" OR "large language models" OR LLM OR LLMs OR GPT* OR Claude OR Gemini OR Llama OR Copilot OR Bard OR DeepSeek OR Qwen)
- **Block 2 (Mathematics):** (math* OR algebra* OR geometr* OR arithmetic OR calculus OR "word problem" OR "word problems" OR "mathematical reasoning")

Nine early scoping seed papers completed Stream 1. Combining those nine seeds with ERIC and PsycNet produced 1,219 raw entries and 1,039 deduplicated core records (272 ERIC-only, 731 APA-only, 27 indexed in both databases, and 9 seed-only).

S1.1.2 Stream 2: Multi-API Scholarly Expansion and Bidirectional Citation Chasing

Stream 2 expanded coverage across six programmatic APIs: OpenAlex (works endpoint, 2022–2026), Crossref (works REST API), Semantic Scholar (graph/v1/paper/search plus citation graphs), the full-window ERIC API, arXiv (cs.CY, cs.AI, cs.CL, stat.AP, math.HO), and EuropePMC. Search clauses paired K–12 schooling descriptors (K-12, primary school, elementary school, middle school, junior high, secondary school, high school, sixth form, Grade 1–Grade 13) with GenAI and mathematics terms, while forward and backward snowballing traced citations around anchor trials (Bastani et al., 2025; Z. Liu et al., 2025; Oreopoulos et al., 2026; R. E. Wang et al., 2025) and prior syntheses (Boolzen et al., 2026; Deng et al., 2025). Table S1.1 reconciles every count.

Table S1.1. *Exact PRISMA 2020 Flow Reconciliation Across Search Streams, Full-Text Retrieval, and Synthesis Strata*

PRISMA 2020 Stage	Component Breakdown	Exact Verified Count
Identification: Stream 1 (Core Register)	ERIC E01–E08 (451) + APA PsycNet (759) + Early Seed (9)	1,219 raw → 1,039 unique
Identification: Stream 2 (Multi-API & Snowball)	OpenAlex (5,532) + Crossref (2,272) + S2/arXiv/ERIC-API/EuropePMC (821) + Snowball (1,217)	9,842 raw → 5,491 net new unique
Deduplication Total	4,531 exact/fuzzy DOI, title-hash, and version duplicates removed across 11,061 raw records	6,530 unique records screened
Title & Abstract Screening	5,940 off-topic/non-K–12/non-GenAI excluded; 218 secondary reviews/meta-analyses catalogued	6,158 excluded at Ti/Ab
Full-Text Retrieval	All 372 sought reports retrieved in full (including all 30 Stream-1 candidate & 14 subscription-access items)	372 retrieved (0 not retrieved)
Full-Text Exclusions (n = 323)	Non-K–12 sample (134); LLM solver/grader benchmark only (79); Teacher-only PD/plans (48); Non-math/STEM/CS (34); Non-GenAI ITS/confounded bundle (12); Conceptual/protocol (11); Statcheck/duplicate diss. (5)	323 excluded with documented reasons
Included in Systematic Review	49 K–12 school-mathematics reports representing 46 independent empirical cohorts	21 quant. reports (20 w/ 36 contrasts + 1 within-AI modality RCT) + 28 qual./single-arm/obs./boundary

Note. Machine-readable counts match prisma_2020_counts.json.

S1.2 Complete Item-by-Item Disposition of Stream-1 Scoping Candidates (n = 30) and Stream-2 Subscription-Access Candidates (n = 14)

https://doi.org/10.1111/bjet.13586Thirty scoping candidates from Stream 1 (R003–R968) required primary-text eligibility and test-condition adjudication, while Stream 2 flagged 14 subscription-access K–12 mathematics articles (PW-01–PW-14). Every report in both ledgers was retrieved and audited in full (Reports not retrieved = 0). Citation hygiene for R170 (Xing et al., 2025) followed strict archival rules: full-text verification cross-checked the complete author manuscript against publisher metadata, and every reference cites the final journal issue (*British Journal of Educational Technology*, 2025, 56(5), 2043–2077, https://doi.org/10.1111/bjet.13586) rather than manuscript pagination.

Table S1.2a. *Complete Full-Text Resolution of All 30 Stream-1 Candidate Scoping Records (R003–R968): 11 Included, 19 Excluded, 0 Unretrieved*

Record ID	Authors & Year	DOI	Final Full-Text Decision	Verified Primary-Text Basis & Synthesis Role
R003	Xuan et al. (2025)	10.1155/hbe2/8886206	INCLUDED (INC-11)	Full publisher PDF verified: Grade 11 limits in 2 Vietnamese schools (2 × N = 60); Tier-2 Cat B contrasts B-06a (+0.457) & B-06b (−1.435).
R010	Rismanchian et al. (2026)	10.48550/arXiv.2605.21629	INCLUDED (INC-09)	Full arXiv v3 PDF verified: 3.2M ALEKS K–12/college learning logs + 12.2M PPL items; Tier-2 Cat A (A-03, +0.340) vs. Proctored Cat D (D-04, −0.159).
R031	Kamau & Shukla (2026)	10.46328/ijtes.5581	EXCLUDED (Non-K–12)	Full PDF verified: both surveys conducted in a public state university gateway mathematics programme; no primary/secondary student data.
R042	Yunianto, Galiç, & Lavicza (2024)	10.46328/ijemst.4437	INCLUDED (INC-46)	Full PDF verified: lower-secondary classroom educational design research integrating ChatGPT + GeoGebra (Cat A qualitative process synthesis).
R061	Alan & Yurt (2024)	10.51383/ijonmes.2024.328	EXCLUDED (Conceptual)	Full PDF verified: theoretical proposal of ‘ChatGPT Aided Flipped Learning Model (CAFLM)’ with illustrative prompts; zero school participants.
R088	Santhi & Srinivasan (2024)	10.1109/TLT.2024.3417252	EXCLUDED (Non-K–12 / CS)	Full IEEE PDF verified: undergraduate engineering cyber-physical attack-detection course using MATLAB; neither

Record ID	Authors & Year	DOI	Final Full-Text Decision	Verified Primary-Text Basis & Synthesis Role
				K–12 nor school mathematics.
R090	Joung & Kim (2026)	10.1111/ssm.18393	EXCLUDED (Non-K–12)	Full Wiley PDF verified: N = 40 undergraduate STEM majors enrolled in a university remedial mathematics course.
R111	Slobodsky, Durcheva, & Kugel (2026)	10.30958/aje.13-1-10	EXCLUDED (Non-K–12)	Full PDF verified: N = 22 first-year teacher-education college students evaluating Chat-Mat™.
R122	Contel & Cusi (2025)	10.1007/s40751-024-00164-7	INCLUDED (INC-47)	Full Springer PDF verified: Italian upper-secondary pilot analysing metacognitive utilisation schemes with GPT-4 during problem solving (Cat A).
R158	Canonigo (2024)	10.1111/jcal.13065	EXCLUDED (Forensic/Bundle)	Full Wiley PDF verified: confounded GeoGebra + ChatGPT arm vs. lecture without GeoGebra, swapped arm Ns across text/tables (30/32 vs. 32/30), incompatible t-tests.
R162	Huang, Kimmins, Winters, & Suh (2025)	10.1108/IJLLS-10-2024-0222	EXCLUDED (Teacher PD)	Full Emerald PDF verified: 4 teachers used ChatGPT during Lesson Study planning; pupils used CODAP/Desmos with no pupil GenAI exposure or tested contrast.
R170	Xing et al. (2025)	10.1111/bjet.13586	INCLUDED (INC-06)	Full manuscript & BJET metadata verified: Study 3 quasi-experiment (N = 212, 4 middle-school classes) on ALTER-Math teachable agent; Tier-1 Cat B (B-02, +0.449).
R174	Canonigo (2026)	10.1002/jcal.70307	INCLUDED (INC-48)	Full Wiley PDF verified: 8-week interpretivist qualitative study in 2 Philippine secondary schools (10 teachers, 24 pupils) on critical mediation & equity (Aux).
R183	Lintner et al. (2026)	10.1007/s10639-026-14002-3	INCLUDED (INC-37)	Full Springer PDF verified: nationally stratified Czech Grade 7 survey (N = 2,307) modelling GenAI use frequency and epistemic offloading vs. IRT math scores (Obs).
R191	Tu, Guo, Fang, Meng, & Hwang (2026)	10.1007/s11423-025-10579-7	EXCLUDED (Non-Math STEAM)	Full ETR&D PDF verified: interdisciplinary STEAM perception mapping without a separable school-

Record ID	Authors & Year	DOI	Final Full-Text Decision	Verified Primary-Text Basis & Synthesis Role
				mathematics achievement outcome.
R308	Adair (2024, Rutgers Diss.)	10.7282/t3-w6vm-m433	EXCLUDED (Automated Grader)	Full dissertation PDF verified: middle-school science inquiry in rule-based Inq-ITS; LLMs tested solely for offline automated scoring of written science explanations.
R309	Adair (2024, Duplicate Entry)	10.7282/t3-w6vm-m433	EXCLUDED (Duplicate Diss.)	Exact repository duplicate of R308 (same DOI and dissertation text).
R325	Luo, Wang, & Hew (2026)	10.1016/j.compedu.2026.105697	INCLUDED (INC-19)	Full Elsevier PDF + Supplement verified: Grade 5 Study 2 (N = 95) Week-7 closed-paper Pólya math exam comparing Teacher HITL+SRL, Auto-SRL, and Unregulated AI (W-04a/b, Cat D).
R339	Aslan, Güney, & Erdoğan (2026)	10.1080/10447318.2026.2678539	EXCLUDED (Preschool)	Full Taylor & Francis PDF verified: N = 76 early-childhood preschool children (ages 4–5) on Concrete patterning; outside K–12 school scope.
R362	Gao et al. (2026)	10.1016/j.neuroimage.2026.122123	EXCLUDED (Non-K–12 / fNIRS)	Full NeuroImage PDF verified: adult university laboratory fNIRS cognitive task comparing LLM vs. web search; neither K–12 nor mathematics curriculum.
R386	Miller (2026)	10.1007/s10758-026-09983-6	EXCLUDED (Non-K–12 / Physics)	Full Springer PDF verified: university introductory physics AI tutor prompt-engineering study; not K–12 school mathematics.
R390	Li, Xing, Li, & Lyu (2026)	10.1016/j.compedu.2026.105725	INCLUDED (INC-21)	Full Elsevier PDF verified: 3-week 2×2×3 voice-design RCT (N = 386 middle-schoolers) inside ALTER-Math; linked to Xing et al. (2025) family (Cat B).
R427	H. Liu et al. (2026)	10.1080/10447318.2026.2616662	INCLUDED (INC-10)	Full IJHCI PDF verified: 300 urban/rural pupils across Grades 5, 8, and 11 (N = 100 each) + eye-tracking; Tier-2 Cat B contrasts B-05a (+0.836), B-05b (+0.637), B-05c (−0.704).
R584	Waqas, Hasan, & Ali (2025)	10.1007/s40593-025-00525-2	EXCLUDED (Non-K–12 / CS)	Full IJAIED PDF verified: university computer science and engineering courses (programming, AI, computer

Record ID	Authors & Year	DOI	Final Full-Text Decision	Verified Primary-Text Basis & Synthesis Role
				architecture).
R811	Korkmaz Güler & Yıldız (2025)	10.1080/00220671.2025.2510387	EXCLUDED (Solver Benchmark)	Full JER PDF verified: researchers inputted 9 static pupil error scripts on exponents into ChatGPT-4 and Gemini to rate diagnostic accuracy; zero pupil AI interaction.
R812	Oğuz Haçat & Kepceoğlu (2025)	10.1080/00220671.2025.2510389	INCLUDED (INC-49)	Full JER PDF verified: exploratory qualitative case study of N = 4 Grade 7 pupils using ChatGPT on geographic/demographic math problems (Cat A).
R833	Torres-Peña et al. (2024)	10.3390/educsci14091019	EXCLUDED (Non-K–12)	Full MDPI PDF verified: undergraduate university calculus course.
R894	D'Ambrosio (2024)	10.7358/ecps-2024-030-ambf	EXCLUDED (Conceptual)	Full Italian PDF verified: prospective description of an unstarted project citing contextual PISA statistics; zero empirical intervention data.
R912	Pardos & Bhandari (2024)	10.1371/journal.pone.0304013	EXCLUDED (Adult MTurk)	Full PLoS ONE PDF verified: N = 274 adult Amazon Mechanical Turk crowdworkers required to hold a high-school diploma.
R968	Wu et al. (2024)	10.1177/07356331231191125	EXCLUDED (Non-K–12 / Non-Math)	Full SAGE PDF verified: university blended learning aid (CILA) without K–12 mathematics outcomes.

Note. Complete full-text PDFs were audited for every record (11 included; 19 excluded for documented PICOS or numerical integrity reasons).

Table S1.2b. *Complete Full-Text Resolution of All 14 Stream-2 Subscription-Access Candidate Articles (PW-01–PW-14): 11 Included, 3 Excluded, 0 Unretrieved*

ID	Authors & Year	DOI	Final Full-Text Decision	Verified Primary-Text Basis & Synthesis Role
PW-01	Kuo, Bai, & Lin (2025)	10.1016/j.compedu.2025.105463	INCLUDED (INC-05)	Grade 5 Socratic + Pólya TALPer companion (Exp 1 N = 363; Exp 2 N = 104); Tier-1 Meta Cat B (B-01a, B-01b).
PW-02	Worden, Vanacore, Haim, & Heffernan (2025)	10.1145/3698205.3729546	INCLUDED (INC-18)	Middle-school ASSISTments RCTs (Exp 1 N = 188; Exp 2 N = 6,371); Within-AI

ID	Authors & Year	DOI	Final Full-Text Decision	Verified Primary-Text Basis & Synthesis Role
				(W-03) & Human Benchmark (H-01).
PW-03	Wang, Prihar, Haim, & Heffernan (2024)	10.1007/978-3-031-64312-5_29	EXCLUDED (Rater Benchmark)	AIED CCIS short paper: educator/crowd ratings of LLM hint quality without student learning outcomes (student RCT reported in PW-02).
PW-04	Zhu et al. (2025)	10.1007/s10639-024-12898-3	EXCLUDED (Non-Math CS)	Full 33-page EAIT PDF verified: Grade 8 Python programming assignments, not mathematics.
PW-05	X. Wang & Wei (2025)	10.1002/acp.70088	INCLUDED (INC-38)	Grade 6 intervention (N = 105) measuring MARS math anxiety and interviews; no math test administered (Aux affective synthesis).
PW-06	Norberg et al. (2025)	10.1007/s40593-024-00420-2	INCLUDED (INC-33)	Full 40-page IJAIED PDF verified (main-text Table 1, Norberg et al., 2025): N = 72,905 pupils in MATHia offline GPT-4 readability rewrites (Cat A/Aux).
PW-07	Serin & Kepceoğlu (2026)	10.1080/00220671.2026.2735824	INCLUDED (INC-40)	Grade 7 classroom study (N = 40, 2 intact classes) comparing GPT-5 Hint vs. Critique feedback in fraction/ratio problem posing (Cat A).
PW-08	Polydoros (2025, IGI Global Ch. 9)	10.4018/979-8-3693-5488-9.ch009	EXCLUDED (Forensic/Bundle)	Single-group pre–post (N = 54 LD pupils) bundling Khan Academy + DreamBox + ChatGPT, with internal timeline (2020 COVID vs. ChatGPT) & grade contradictions.
PW-09	Ergene & Çaylan Ergene (2025)	10.1007/s10639-024-13121-z	INCLUDED (INC-39)	160 high-schoolers & 80 math teachers evaluating 5 GenAI chatbots across 390 e-textbook tasks (Aux epistemic evaluation synthesis).
PW-10	Dello Iacono & Picariello (2026)	10.1564/tme_v33.2.05	INCLUDED (INC-41)	Grade 13 Italian class (N = 12) comparing unstructured vs. structured tutor prompt on calculus derivatives metacognition (Cat A).
PW-11	Y. Kim, Lee, Shin, & Park (2025)	10.57090/sm.2025.12.27.4.921	INCLUDED (INC-42)	Primary-school commognitive discourse case study on arithmetic mean levelling-off with GenAI self-explanation prompts (Cat A).
PW-12	N. Kim & Chang (2026)	10.29275/jerm.2026.36.3.461	INCLUDED (INC-43)	Grade 5 qualitative case study (N = 6) of

ID	Authors & Year	DOI	Final Full-Text Decision	Verified Primary-Text Basis & Synthesis Role
				real-world context integration in mathematical modelling with GenAI (Cat A).
PW-13	So, Jung, Seong, & Kim (2026)	10.29275/jerm.2026.36.3.401	INCLUDED (INC-44)	Grade 10 commognitive case study (N = 3) of roller-coaster curve modelling before and during ChatGPT mediation (Cat A).
PW-14	Lee, Park, Shin, & Kwak (2025)	10.30807/ksms.2025.28.4.007	INCLUDED (INC-45)	Grade 9 qualitative study (N = 6) of Pfannkuch statistical reasoning & counter-arguments on box plots with AI chatbot (Cat A).

Note. Combined with Table S1.2a, primary full texts were verified for 100.0% of subscription-access and Stream-1 K–12 candidates.

S1.3 Forensic Numerical Audit, Cohort Linkage, and Integrity-Gate Exclusions

Independent scripts checked numerical consistency across every quantitative study. Automated statcheck routines recomputed *t, F, df, p,* pooled SDs, pretest–posttest difference-in-differences, and raw repository microdata wherever available. Table S1.3 records each forensic resolution, cross-links companion cohorts to prevent double-counting, and details the five full-text integrity exclusions triggered by irreconcilable arithmetic or co-intervention flaws (four statcheck exclusions grouped with duplicate dissertation R309 inside the *n* = 5 PRISMA integrity/duplicate gate in Table S1.1, plus uncontrolled multi-platform bundle PW-08 inside the *n* = 12 confounded-bundle gate).

Table S1.3. *Forensic Numerical Audit Findings, Companion-Cohort Linkages, and Full-Text Integrity Exclusions*

Source & Identifier	Audit Disposition	Exact Forensic Verification Finding
Bastani et al. (2025), PNAS + GitHub CSV	Included (Tier-1, Cat A & D, Within-AI W-01)	Raw CSV audit confirmed 825 unique pupils across 44 non-honours classes (2,848 pupil–session rows), yielded empirical classroom ICCs (ρ = 0.361 in Part 2; ρ = 0.128 in Part 3), and directly resolved the unpublished head-to-head Part 3 closed-exam contrast (GPT Tutor vs. GPT Base: β = +0.0500, SE_cl = 0.0233, t = 2.15, p = .0377, g = +0.188).
Kuo, Bai, & Lin (2025), C&E	Included (Tier-1, Cat B)	Text states Group 1 N = 175 while all ANCOVA/mean tables report N = 174 (verified N = 174 used); teachers chose experimental arms across 8 schools; in Exp 2 (N = 104), unequal low-achiever proportions (67% in G5 vs. 32% in G4) create a raw-mean Simpson’s paradox resolved via Johnson–Neyman

Source & Identifier	Audit Disposition	Exact Forensic Verification Finding
		subgroup stratification (LA +15.06 pts, p = .017).
Cosentino et al. (2025), BJET	Included (Active Human Benchmark H-02)	Section 4 states ‘no significant difference (t(31.63) = 0.73, p < 0.05)’: exact two-sided p for |t| = 0.73 (df = 31.63) is p = .4708 (typographical ‘<’ printed for ‘>’); 30 of 34 enrolled pupils analysed (n = 14 AI vs. n = 16 teacher).
J. Liu et al. (2025a CAEAI; 2025b GCCCE; 2025c SLE)	Included as Linked ChatGPT-MPS Cohort (Not Pooled)	CAEAI journal article and Chinese GCCCE proceedings report identical paired 12-item statistics for n = 49 retained AI-arm pupils while omitting posttest scores for the comparison class; linked without double-counting.
Xing et al. (2025 BJET) + Z. Liu et al. (2025 SciRep) + Li et al. (2026 C&E)	Included as Linked ALTER-Math Programme	Only Study 3 of Xing et al. (2025; N = 212, 4 classes) enters GenAI-vs.-non-AI pooling (B-02). Z. Liu et al. (2025; N = 589 single-group pre–post where 327 improved and 206 declined) and Li et al. (2026; N = 386 voice-design RCT across AI arms) are synthesised narratively without overlap.
Xuan et al. (2025), HBET	Included (Tier-2 Sensitivity, B-06a/b)	Full text revealed opposite-signed school effects (+0.457 in IT-rich FPT private school vs. −1.435 in under-resourced Thoi Thuan public school) despite significantly higher baseline IQ in the public-school AI arm (98.3 vs. 91.4, p < .01).
H. Liu et al. (2026), IJHCI	Included (Tier-2 Sensitivity, B-05a/b/c)	Kept in Tier 2 because Methods text alternates between describing the comparator as ‘traditional classroom teaching’ and ‘independent search-and-read’ and alternates between 3-week and 8-week durations.
Davor (2026), Front. Educ.	Included (Tier-2 Sensitivity, B-10)	Pretest group means swap labels between Table 3 (EG 13.21 vs. CG 12.37) and Table 4 (CG 13.21 vs. EG 12.37) in a 2-class intact design; posttest SMD with K = 2 cluster inflation retained strictly in Tier-2 sensitivity.
Nguyen et al. (2026), Educ. Sci.	Included as Flagged Boundary Report (Not Pooled)	Teacher-delivered cultural storytelling geometry videos (no live pupil AI interaction) plus internal discrepancy between reported independent-samples t-test and printed posttest SDs.
Chen (2025), ACM AIFE (10.1145/3785987.3786040)	EXCLUDED at Full Text (Integrity Gate)	Printed posttest M/SD across 4 arms (n = 60 each: 79.5±7.1, 82.7±6.3, 81.2±5.9, 64.3±7.8) mathematically require one-way ANOVA F(3, 236) = 123.70, contradicting printed F = 47.32; additionally treats Stability AI image generator as a math tutor and cites zero post-2020 sources.
Canonigo (2024), JCAL (10.1111/jcal.13065, R158)	EXCLUDED at Full Text (Integrity/Bundle Gate)	Experimental arm bundled GeoGebra dynamic graphing with ChatGPT against lecture without GeoGebra, swapped group sample sizes across text and tables (30/32 vs. 32/30), and reported incompatible paired/independent t-values.

Source & Identifier	Audit Disposition	Exact Forensic Verification Finding
Polydoros et al. (2025), Grade 6 Geometry (N = 436)	EXCLUDED at Full Text (Integrity Gate)	Statcheck recalculation from printed group means and standard deviations contradicted reported test statistics (t = 8.47 impossibility).
Polydoros (2025), IGI Global Ch. 9 (PW-08)	EXCLUDED at Full Text (Bundle/Timeline Gate)	Single-group pre–post without control (N = 54 LD pupils), inseparable Khan Academy + DreamBox + ChatGPT bundle, and internal contradictions (ChatGPT use attributed to COVID-19 lockdown period; Grade 5 vs. Grade 6; 4 vs. 8 weeks).
Nangolo (2026), Non-indexed Preprint/Article	EXCLUDED at Full Text (Integrity Gate)	Unverifiable synthetic variance structure failing statcheck consistency.

Note. Every exclusion and sensitivity rule executes reproducibly in run_final_meta_analysis.py.

S1.4 Complete Effect-Size Derivation Ledger Across All 36 Extracted Contrasts

Exact arm denominators, cluster counts, raw or covariate-adjusted test statistics, intraclass correlation (ICC) design effects, unclustered Hedges' *g* (SE), and cluster-adjusted *g*_adj (SE_adj) appear in Table S1.4 for all 36 contrasts stored in All_36_Meta_Analytic_Contrasts.csv (14 Tier-1 Primary GenAI-vs.-Non-AI, 14 Tier-2 Sensitivity GenAI-vs.-Non-AI, 2 Active Expert Human Benchmarks, and 6 Direct Head-to-Head Within-AI Guardrail Contrasts).

Table S1.4. *Complete Effect-Size and Cluster-Variance Derivation Ledger for All 36 School-Mathematics Contrasts*

ID	Study [Cat / Ped.]	Sample & Cluster Structure	Extracted Arm Outcomes & Test Statistic	Unclust. g (SE)	Cluster-Adj g (SE)	Pool
A-01a	Bastani et al. (2025) [A / M1]	T=242, C=320 (30 classes: 13 T / 17 C; shared C split n_C/2)	T: 0.474 (0.258) vs. C: 0.284 (0.287); beta = +0.137 (SE_cl = 0.031), t = 4.42, p < .0001	+0.485 (0.110)	+0.485 (0.131)	Prim
A-01b	Bastani et al. (2025) [A / M2]	T=277, C=320 (31 classes: 14 T / 17 C; shared C split n_C/2)	T: 0.669 (0.309) vs. C: 0.284 (0.287); beta = +0.361 (SE_cl = 0.032), t = 11.38, p < .0001	+1.181 (0.104)	+1.181 (0.125)	Prim
A-02	Oreopoulos et al. (2026) [A / M3]	T=1750, C=1820 (Indiv. login randomisation; N = 3,570 mastery stratum)	T: 45.0% Q1 first-attempt correct vs. C: 36.2%; Delta_p = +0.088 (SE = 0.016), z = 5.50, p < .001	+0.202 (0.036)	+0.202 (0.036)	Prim
A-03	Rismanchian et al. (2026) [A / M1]	3.2M learning + 12.2M PPL interactions (66 topics FE)	T: OR = 1.85 (+85% unproctored odds) vs. C: AI-resistant graph baseline; beta_logodds = +0.0561/qtr (+0.617 cum.), p < .001	+0.340 (0.061)	+0.340 (0.061)	Ext
B-01a	Kuo et al. (2025) [B / M3]	T=189, C=174 (16 classes across 8 schools; teacher-chosen arm)	T: Post 74.76 (16.52), Adj 76.19 (1.13) vs. C: Post 73.27 (20.95), Adj 71.69 (1.17); ANCOVA F = 7.635, p = .001 (G3 vs G1 +4.50, p = .018)	+0.239 (0.086)	+0.238 (0.205)	Prim

ID	Study [Cat / Ped.]	Sample & Cluster Structure	Extracted Arm Outcomes & Test Statistic	Unclust. g (SE)	Cluster-Adj g (SE)	Pool
B-01b	Kuo et al. (2025) [B / M3]	T=60, C=44 (9 classes, 9 teachers; Johnson-Neyman stratified)	T: HA 89.55 (7.74) [n=20], LA 72.90 (20.61) [n=40] vs. C: HA 89.30 (9.43) [n=30], LA 57.71 (27.19) [n=14]; LA ANCOVA F = 6.142, p = .017 (+15.06 pts); HA F = 0.202, p = .655	+0.392 (0.189)	+0.387 (0.307)	Prim
B-02	Xing et al. (2025) [B / M3]	T=107, C=105 (4 classes, 4 teachers; Math Nation platform)	T: Pre 0.32 (0.14) -> Post 0.43 (0.20) vs. C: Pre 0.35 (0.16) -> Post 0.39 (0.19); OLS group b = +0.08 [0.02, 0.14], p = .009 (unclustered)	+0.464 (0.137)	+0.449 (0.385)	Prim
B-03	Tasdelen & Bodemer (2025) [B / M3]	T=57, C=57 (Indiv. randomisation within classrooms)	T: 4.97 (2.40) / 10 vs. C: 4.04 (2.67) / 10; F(1,112) = 3.82, two-sided p = .053 (one-sided p = .027)	+0.364 (0.188)	+0.364 (0.188)	Prim
B-04	Lademann et al. (2025) [B / M2]	T=107, C=107 (Indiv. randomisation)	T: 50.29% (guesses = 0) vs. C: 50.38% (guesses = 0); Mann-Whitney p = .75 (null)	-0.004 (0.136)	-0.004 (0.136)	Prim
B-05a	Liu, H. et al. (2026) [B / M2]	T=50, C=50 (4 regional cells/classes, 25 per cell; Grade 5)	T: 96.19 (2.31) vs. C: 93.64 (3.46); Post diff +2.56 pts, g_indiv = +0.864	+0.864 (0.208)	+0.836 (0.421)	Ext
B-05b	Liu, H. et al. (2026) [B / M2]	T=50, C=50 (4 regional cells/classes, 25 per cell; Grade 8)	T: 89.67 (6.08) vs. C: 85.58 (6.26); Post diff +4.09 pts, g_indiv = +0.657	+0.657 (0.204)	+0.637 (0.414)	Ext
B-05c	Liu, H. et al. (2026) [B / M2]	T=50, C=50 (4 regional cells/classes, 25 per cell; Grade 11)	T: 65.03 (14.44) vs. C: 74.46 (11.09); Post diff -9.43 pts, g_indiv = -0.727	-0.727 (0.205)	-0.704 (0.416)	Ext
B-06a	Xuan et al. (2025) [B / M1]	T=30, C=30 (2 groups: 30 T / 30 C; IT-rich private school)	T: MCQ Post 4.99 (gain +3.83) vs. C: 4.58 (gain +3.18); t(58) = +2.28, p = .026; OLS beta_AI = +0.785 (SE = 0.409)	+0.489 (0.259)	+0.457 (0.568)	Ext
B-06b	Xuan et al. (2025) [B / M1]	T=30, C=30 (2 groups: 30 T / 30 C; under-resourced public school)	T: MCQ Post 4.58 (gain +3.28) vs. C: 6.61 (gain +5.08); t(58) = -9.12, p < .001; OLS beta_AI = -2.125 (SE = 0.352)	-1.536 (0.291)	-1.435 (0.628)	Ext
B-07	Lin et al. (2025) [B / M3]	T=28, C=31 (2 intact classes: 1 T / 1 C)	T: Pre 3.61 (1.55) -> Post 5.71 (2.21) vs. C: Pre 3.23 (1.50) -> Post 4.58 (1.79); Mann-Whitney U = 316, p = .069 (Low-achievers p = .010)	+0.558 (0.262)	+0.454 (0.569)	Ext
B-08	Karaman & Göksu (2024) [B / M2]	T=24, C=15 (2 intact classes: 1 T / 1 C)	T: Pre 38.42 (11.95) -> Post 61.21 (22.43) vs. C: Pre 34.00 (14.72) -> Post 55.27 (19.03); Post t(37) = 0.851, p = .400 (n.s.)	+0.274 (0.324)	+0.106 (0.604)	Ext
B-09	Malik et al. (2026) [B / M3]	T=24, C=23 (2 intact classes: 1 T / 1 C)	T: Pre 53.33 (15.23) -> Post 68.33 (12.74) vs. C: Pre 49.13 (12.40) -> Post 59.57 (11.86); Post t(45) = 2.439, p = .019 (+4.20 baseline imbalance)	+0.699 (0.296)	+0.301 (0.578)	Ext
B-10	Davor (2026) [B / M3]	T=43, C=43 (2 intact classes: 1 T / 1 C)	T: Pre 13.21/12.37* -> Post 21.74 (4.42) vs. C: Pre 12.37/13.21* -> Post 16.79 (4.76); Post t(84) = 5.00, ANCOVA F(1,83) = 50.30, p < .001	+1.068 (0.229)	+0.997 (0.577)	Ext

ID	Study [Cat / Ped.]	Sample & Cluster Structure	Extracted Arm Outcomes & Test Statistic	Unclust. g (SE)	Cluster-Adj g (SE)	Pool
B-11	Prabhakar (2026) [B / M3]	T=50, C=50 (Intact classrooms, est. K = 4)	T: Pre 35.2 (8.6) -> Post 48.7 (7.9) vs. C: Pre 34.8 (8.9) -> Post 38.4 (8.7); ANCOVA F(1,95) = 63.87, p < .001	+1.230 (0.217)	+1.087 (0.432)	Ext
B-12a	Polydoros (2026) [B / M1]	T=132, C=132 (9 schools; baseline school ICC = .03; shared C split)	T: 69.62 (10.74) vs. C: 67.84 (10.92); Post diff +1.78, p > .05 (n.s.)	+0.164 (0.123)	+0.164 (0.203)	Ext
B-12b	Polydoros (2026) [B / M3]	T=132, C=132 (9 schools; baseline school ICC = .03; shared C split)	T: 76.48 (9.86) vs. C: 67.84 (10.92); Post diff +8.64, p < .001	+0.828 (0.128)	+0.828 (0.212)	Ext
C-01	Wang et al. (2025) [C / M3]	T=500, C=500 (900 tutors; cluster-robust SE by tutor-pupil pair)	T: 65.7% unconditional pass (76.3% cond.) vs. C: 61.7% (73.2% cond.); Delta_p = +0.040 (SE = 0.015), Romano-Wolf p = .010	+0.095 (0.037)	+0.095 (0.037)	Prim
D-01a	Bastani et al. (2025) [D / M1]	T=242, C=320 (30 classes: 13 T / 17 C; shared C split n_C/2)	T: 0.287 (0.233) vs. C: 0.321 (0.277); beta = -0.0544 (SE_cl = 0.0224), t = -2.42, p = .020	-0.205 (0.084)	-0.205 (0.101)	Prim
D-01b	Bastani et al. (2025) [D / M2]	T=277, C=320 (31 classes: 14 T / 17 C; shared C split n_C/2)	T: 0.315 (0.281) vs. C: 0.321 (0.277); beta = -0.0043 (SE_cl = 0.0134), t = -0.32, p = .749	-0.015 (0.047)	-0.015 (0.057)	Prim
D-02a	Oreopoulos et al. (2026) [D / M2]	T=1525, C=1593 (Indiv. login RCT; N = 3,118 non-mastery stratum)	T: 0.739 / 2 practiced items vs. C: 0.780 / 2; beta = -0.041 (SE = 0.028), p = .143 (DID p = .68)	-0.048 (0.033)	-0.048 (0.033)	Prim
D-02b	Oreopoulos et al. (2026) [D / M3]	T=1569, C=1640 (Indiv. login RCT; N = 3,209 mastery stratum)	T: 40.2% Ex.1 correct (Hard: 49.9%) vs. C: 37.0% (Hard: 44.0%); Delta_p = +0.032 (SE = 0.017), p = .065 (Hard +0.059, p = .017)	+0.074 (0.039)	+0.074 (0.039)	Prim
D-03	Henkel et al. (2024) [D / M3]	T=237, C=241 (11 schools; school-level assignment)	T: Pre 20.29 (8.72) -> End 25.42 (7.25) vs. C: Pre 20.20 (8.81) -> End 22.32 (8.06); DID +3.01 pts, reported d = 0.36 (unclustered p < .001)	+0.360 (0.087)	+0.356 (0.234)	Prim
D-04	Rismanchian et al. (2026) [D / M1]	6,698 randomly triggered proctored retention items	T: OR = 0.75 (-25% retention odds), 95% CI [0.56, 0.98]; beta_logodds = -0.288 cum., permutation p < .05	-0.159 (0.079)	-0.159 (0.079)	Ext
H-01	Worden et al. (2025) [C / M2]	T=4523, C=4246 (6,371 pupils, 24,378 help requests; student + item RE)	T: Log-odds diff +0.01 (rest of set +0.00) vs. C: Expert teacher-written explanation; beta_logodds = +0.01 [-0.06, 0.08], p = .740	+0.006 (0.020)	+0.006 (0.020)	Active Human Benchmark
H-02	Cosentino et al. (2025) [A / M2]	T=14, C=16 (Indiv. randomisation; 34 enrolled, 30 analysed)	T: Session accuracy ratio (n = 14) vs. C: Clinical teacher feedback (n = 16); |t| = 0.73, two-sided p = .471 (misprinted 'p < 0.05' in text)	-0.260 (0.358)	-0.260 (0.358)	Active Human Benchmark
W-01	Bastani et al. (2025) [D / Guardrail]	T=277, C=242 (27 classes: 14 Tutor / 13 Base)	T: 0.315 (0.281) vs. C: 0.287 (0.233); beta = +0.0500 (SE_cl = 0.0233), t = 2.15, p = .038	+0.188 (0.087)	+0.188 (0.087)	Within AI Guardrail
W-02	Oreopoulos et al. (2026) [D / Mastery]	T=1569, C=1525 (Indiv. login RCT factorial contrast)	T: +3.2 pp AI gain under Mastery vs. C: -2.1 pp AI loss without Mastery; Interaction Delta_g = +0.123 (SE = 0.051), z = 2.41, p = .016	+0.122 (0.051)	+0.122 (0.051)	Within AI Guardrail

ID	Study [Cat / Ped.]	Sample & Cluster Structure	Extracted Arm Outcomes & Test Statistic	Unclust. g (SE)	Cluster-Adj g (SE)	Pool
W-03	Worden et al. (2025) [C / Explanation]	T=125, C=139 (188 pupils, 1,314 help requests; student + item RE)	T: OR = 4.18 on next problem (+2 pp rest of set) vs. C: Direct answer without explanation; beta_logodds = +1.43 [0.67, 2.18], p < .001	+0.788 (0.212)	+0.788 (0.212)	Within AI Guardrail
W-04a	Luo et al. (2026) [D / SRL]	T=33, C=32 (Grade 5 cohort; shared CG split)	T: Median 8.0 (IQR 7.0) / 40 vs. C: Median 4.0 (IQR 4.0) / 40; Mann-Whitney Z = 3.185, p_adj = .002, r = 0.40 (AUC = 0.728)	+0.847 (0.256)	+0.847 (0.315)	Within AI Guardrail
W-04b	Luo et al. (2026) [D / TeacherHITL]	T=30, C=32 (Grade 5 cohort; shared CG split)	T: Median 13.0 (IQR 6.0) / 40 vs. C: Median 4.0 (IQR 4.0) / 40; Mann-Whitney Z = 5.225, p_adj < .001, r = 0.66 (AUC = 0.885)	+1.676 (0.293)	+1.676 (0.356)	Within AI Guardrail
W-05	Polydoros (2026) [B / TeacherMediated]	T=132, C=132 (Intact classes across 9 schools; ICC = .03)	T: 76.48 (9.86) vs. C: 69.62 (10.74); Post diff +6.86 pts, p < .001	+0.664 (0.126)	+0.664 (0.170)	Within AI Guardrail

Note. Prim = Model 1 Primary Tier-1 and Model 2 Extended GenAI-vs.-Non-AI; Ext = Model 2 Extended Sensitivity; Within AI Guardrail = Model 3c; Active Human Benchmark = Model 3d. Cluster variance adjustments apply empirical classroom ICC ρ = 0.128 whenever primary reports gave unclustered standard errors. Pairwise two-arm pooled SDs and exact degrees-of-freedom *J* corrections govern head-to-head contrasts in Model 3c (main-text Table 3), which differ by ≤ 0.005 from three-arm ledger values. *In B-10, Davor (2026) reverses pretest group labels between Table 3 and Table 4.

S1.5 Result-Specific Methodological Quality Appraisals Across All 49 Included Reports

Critical appraisal operated at the individual contrast level rather than through blanket paper scores. Evaluative frameworks matched each study design: Cochrane RoB 2 for individual RCTs (D1 randomization, D2 deviations from intended interventions, D3 missing outcome data, D4 outcome measurement, D5 selection of reported result), the RoB 2 cluster extension (adding D1b recruitment timing), and Joanna Briggs Institute (JBI) checklists for quasi-experimental (9 items; JBI, 2024a), analytical cross-sectional (8-item manual dated 26 March 2024; JBI, 2024b), and qualitative (10 items; JBI, 2025) studies.

Table S1.5. *Result-Specific Critical Appraisals Across All 49 Included School-Mathematics Reports (RoB 2 and JBI Frameworks)*

Included Report	Appraisal Tool	Appraised Result / Component Target	Domain Summary & Key Interpretational Limit
INC-01 Bastani et al. (2025)	RoB 2 Cluster RCT	Closed-resource Part 3 exam (D-01a/b) & Part 2 practice (A-01a/b)	Some concerns (D1–D1b Low; D2 Some concerns: open-label classroom AI; D3–D5 Low; preregistered AsPredicted #144093 + open CSV).
INC-02 Oreopoulos et al. (2026)	RoB 2 Individual RCT	Week-2 AI-disabled silent test (D-02a/b) & Week-1 practice (A-02)	Low to Some concerns (D1 Low: login-level randomisation; D2 Low; D3 Some concerns: classroom attendance attrition ~14% balanced across arms; D4–D5 Low).

Included Report	Appraisal Tool	Appraised Result / Component Target	Domain Summary & Key Interpretational Limit
INC-03 Henkel et al. (2024)	JBI Quasi-Exp. / Cluster	8-month supervised paper endline (D-03, 11 schools)	Moderate/Serious concerns (school-level clustering without reported ICC; Rori bundled with +1 hr/wk extra math study hall; paper test blind to AI access).
INC-04 Wang et al. (2025)	RoB 2 Cluster RCT	Unconditional session exit-ticket pass under no-tutor-help rule (C-01)	Some concerns (D1–D1b Low: 900 tutors randomised, OSF 8d6ha; D2 Low; D3 Some concerns: unattempted exit tickets missing in conditional model, resolved via unconditional analysis; D4 Some concerns: remote pupil resources unmonitored).
INC-05 Kuo, Bai, & Lin (2025)	JBI Quasi-Experimental	Immediate Ratios/Units posttest in Exp 1 (B-01a) & Exp 2 (B-01b)	Moderate/High risk (teachers selected assigned arm across 8 schools; test invigilation rules undocumented; minor N = 175 vs. 174 table discrepancy; strong ANCOVA & Johnson–Neyman reporting).
INC-06 Xing et al. (2025)	JBI Quasi-Experimental	Study 3 15-item Math Nation posttest (B-02, 4 classes)	Moderate risk (4 intact classes across 4 teachers without cluster SEs; identical 15-item pre/post instrument; undocumented test-time resource lock-down).
INC-07 Tasdelen & Bodemer (2025)	RoB 2 Individual RCT	Immediate 10-item word-problem posttest (B-03)	Some concerns (D1 Low; D2–D3 Low; D4 Some concerns: immediate classroom posttest with undocumented resource rules; D5 Some concerns: one-sided $p = .027$ reported alongside $F = 3.82$, two-sided $p = .053$).
INC-08 Lademann et al. (2025)	RoB 2 Individual RCT	Confidence-filtered 4-task proportional reasoning test (B-04)	Some concerns (D1–D3 Low; D4 Low/Some concerns: guesses scored zero, exact per-arm Ns & raw SDs omitted from main text; D5 Low).
INC-09 Rismanchian et al. (2026)	JBI Quasi-Exp. (DiD)	ALEKS unproctored ramp (A-03) vs. proctored retention (D-04)	Moderate risk (strong 2-way topic × quarter fixed-effects falsification across 15.4M logs, but individual pupil AI use unobserved and PPL retention sample concentrated in gateway college placement).
INC-10 Liu, H. et al. (2026)	JBI Quasi-Experimental	Grades 5, 8, 11 unit posttests (B-05a/b/c)	High risk (comparator description alternates between traditional teaching and self-study search-and-read; duration alternates between 3 and 8 weeks; pretest baseline unadjusted).
INC-11 Xuan et al. (2025)	JBI Quasi-Experimental	Grade 11 limits posttests in FPT Private & Thoi Thuan Public (B-06a/b)	High risk (baseline IQ imbalance $p < .01$; unverified test-time rules; 2 school cohorts with opposite-signed unclustered effects).
INC-12 Lin et al. (2025)	JBI Quasi-Experimental	Grade 5 authentic problem-solving posttest (B-07, 2 classes)	High risk for causal between-group inference (1 intact class per arm, confounding treatment with classroom teacher; undocumented posttest proctoring).

Included Report	Appraisal Tool	Appraised Result / Component Target	Domain Summary & Key Interpretational Limit
INC-13 Karaman & Göksu (2024)	JBI Quasi-Experimental	Grade 3 5-week math posttest (B-08, 2 classes)	High risk (1 intact class per arm, N = 24 vs. 15; teacher-only ChatGPT lesson planning; unclustered t-test).
INC-14 Malik et al. (2026)	JBI Quasi-Experimental	Grade 8 linear systems posttest (B-09, 2 classes)	High risk (1 intact class per arm, N = 47; +4.20 baseline imbalance; undocumented test proctoring).
INC-15 Davor (2026)	JBI Quasi-Experimental	Senior-high geometry posttest (B-10, 2 classes)	High risk (1 intact class per arm; pretest means swap labels between Tables 3 and 4; undocumented test-time resource closure).
INC-16 Prabhakar (2026)	JBI Quasi-Experimental	Grade 7 mathematics posttest (B-11, N = 100)	High risk (non-indexed venue; intact classes without reported K or ICC; abstract mislabels within-group d_z = 1.62 as between-group effect).
INC-17 Polydoros (2026, Preprint)	JBI Quasi-Experimental	Grade 7 algebra 3-arm posttest (B-12a/b, W-05)	High/Unclear risk (unreviewed preprint; intact classes across 9 schools; author has separate statcheck exclusion in 2025 geometry paper).
INC-18 Worden et al. (2025)	RoB 2 Individual/Attempt RCT	Unassisted next-problem correctness in Exp 1 (W-03) & Exp 2 (H-01)	Low risk (randomised within ASSISTments platform across N = 188 and N = 6,371 pupils; crossed student/item random effects; objective log outcome).
INC-19 Luo, Wang, & Hew (2026)	JBI Quasi-Exp. / 3-Arm Trial	Week-7 closed-paper Pólya word-problem exam (W-04a/b)	Low-to-Moderate risk (explicitly closed-paper unaided exam scored blind with high inter-rater κ; non-normal score distribution appropriately analysed via Holm-adjusted Mann–Whitney tests; modest arm sizes n = 30–33).
INC-20 Cosentino et al. (2025)	RoB 2 Individual RCT	Concurrent embodied number-line session performance (H-02)	Some concerns/High imprecision (small N = 30 analysed of 34 enrolled; typographical 'p < 0.05' misprint for null t = 0.73; concurrent supported performance only).
INC-21 Li, Xing et al. (2026)	RoB 2 Individual/Class RCT	3-week 2×2×3 voice-design posttest inside ALTER-Math (N = 386)	Some concerns (well-powered factorial voice-modality comparison within GenAI teachable agent; no non-AI control arm; immediate platform test).
INC-22 Liu, J. et al. (2025a, CAEAI)	JBI Quasi-Experimental	Paired 12-item score for n = 49 retained ChatGPT-MPS pupils	High risk (comparison-class posttest omitted; selective retention; linked to INC-23).
INC-23 Liu, J. et al. (2025b,	JBI Quasi-Experimental	Paired 12-item score for n = 49 pupils (Chinese proceedings)	High risk (exact cohort overlap with INC-22; single-arm change only).

Included Report	Appraisal Tool	Appraised Result / Component Target	Domain Summary & Key Interpretational Limit
GCCCE)			
INC-24 Liu, J. et al. (2025c, SLE)	JBI Analytical Cross-Sect.	Perceived engagement & usability survey + 3 teacher interviews	Moderate quality for perceptions; does not measure mathematics attainment.
INC-25 Liu, Z. et al. (2025, SciRep)	JBI Quasi-Exp. & Cross-Sect.	Single-group pre–post (N = 589) & gain-selected log strata	High risk for causal inference (no control group; gain strata defined post hoc on identical 15 items where 35% declined).
INC-26 Öner (2025)	JBI Quasi-Experimental	Single-group Grade 6 arithmetic-mean pre–post (N = 13)	High risk (no control group; N = 13; single classroom activity).
INC-27 Bulut & Borromeo Ferri (2026)	JBI Qualitative	Observed pair modelling & verification with ChatGPT	Good qualitative credibility (triangulated video, chat logs, and interviews; supported process only).
INC-28 Rizos et al. (2024)	JBI Qualitative	2 focal Grade 9 SEN pupils using ChatGPT worksheets	Moderate qualitative case depth (illustrative classroom episodes; non-generalisable N = 2 focal case).
INC-29 Walkington et al. (2025)	JBI Qualitative	36 middle-school girls posing algebra problems with AI in camp	Strong qualitative interaction coding; informal camp setting without post-withdrawal test.
INC-30 Besas, Patac, & Patac (2026)	JBI Cross-Sect. & Qual.	Self-directed ChatGPT use vs. senior-high math grades	Moderate/High confounding risk (self-reported frequency; unadjusted cross-sectional association).
INC-31 Nguyen et al. (2026)	JBI Quasi-Experimental	Primary geometry cultural-storytelling AI media (2 classes)	High risk (1 class per arm; teacher-delivered video rather than pupil AI use; internal t/SD statcheck discrepancy).
INC-32 Trocado et al. (2026)	JBI Qualitative	13 Grade 10 pairs exploring quadratic functions with ChatGPT	Moderate qualitative credibility (authentic classroom artefacts; no unaided posttest).
INC-33 Norberg et al. (2025)	RoB 2 Embedded RCT	MATHia workspace promotion under offline GPT-4 rewrites (N = 72,905)	Low internal RoB for embedded A/B text simplification; boundary intervention (pupils never interacted with live GenAI).
INC-34 Baez et al. (2025)	JBI Analytical Cross-Sect.	Educator ratings of 12 AI-assisted polynomial art artefacts	High selection/measurement risk (ratings of curated artefacts, not student mathematics learning).
INC-35 Kuzu (2026)	JBI Qualitative	Multilingual primary small-group translanguaging with ChatGPT	Good interpretative video/transcript rigour; concurrent supported interaction only.

Included Report	Appraisal Tool	Appraised Result / Component Target	Domain Summary & Key Interpretational Limit
INC-36 Karatayeva et al. (2026)	JBI Quasi-Experimental	Grade 4 static Gemini 2.5 Pro paper handouts (N = 263, 10 classes)	Moderate risk (matched within-school classes sharing 5 teachers; multilevel ICC modelled; static paper boundary without live pupil AI).
INC-37 Lintner et al. (2026)	JBI Analytical Cross-Sect.	Czech Grade 7 GenAI use & offloading vs. IRT math (N = 2,307)	High measurement validity (nationally stratified sample, IRT math test, SEM controls), observational cross-sectional design.
INC-38 Wang, X., & Wei, Y. (2025)	JBI Quasi-Experimental	Grade 6 math anxiety MARS pre–post (N = 105, 2 classes)	Moderate/High risk (intact classes; self-report anxiety outcome without objective math test).
INC-39 Ergene & Çaylan Ergene (2025)	JBI Cross-Sect. & Qual.	160 high-schoolers & 80 teachers evaluating 5 AI solvers	Good descriptive/instrumental rigour across 390 problems; evaluates solver affordances rather than post-withdrawal student learning.
INC-40 Serin & Kepceoğlu (2026)	JBI Quasi-Exp. / Process	Grade 7 GPT-5 Hint vs. Critique problem-posing uptake (N = 40)	Moderate process rigour (244 turns coded; authors explicitly note class, topic, and session order are not separable for causal claims).
INC-41 Dello Iacono & Picariello (2026)	JBI Qualitative	Grade 13 calculus derivatives metacognitive coding (N = 12)	Good qualitative taxonomy coding (Meijer framework across two prompt sessions; single humanities class, no unaided test).
INC-42 Kim, Lee, Shin, & Park (2025)	JBI Qualitative	Primary arithmetic-mean commognitive discourse case study	Good commognitive transcript & visual-mediator analysis; small purposeful case without post-withdrawal test.
INC-43 Kim & Chang (2026)	JBI Qualitative	Grade 5 mathematical modelling context integration (N = 6)	Strong multi-source qualitative triangulation (logs, worksheets, reflections, interviews) in researcher's homeroom after-school group.
INC-44 So, Jung, Seong, & Kim (2026)	JBI Qualitative	Grade 10 commognitive roller-coaster modelling (N = 3)	Detailed turn-by-turn commognitive coding across pre-ChatGPT and ChatGPT phases; N = 3 above-average volunteers.
INC-45 Lee, Park, Shin, & Kwak (2025)	JBI Qualitative	Grade 9 box-plot statistical reasoning with AI counter-arguments (N = 6)	Rigorous Pfannkuch reasoning-element coding; supported dialogic phase only.
INC-46	JBI Qualitative	Lower-secondary ChatGPT + GeoGebra computational thinking	Moderate EDR process documentation (9 of 16 volunteer

Included Report	Appraisal Tool	Appraised Result / Component Target	Domain Summary & Key Interpretational Limit
Yunianto, Galiç, & Lavicza (2024)		EDR	chat logs analysed; no unaided geometry transfer test).
INC-47 Contel & Cusi (2025)	JBI Qualitative	Upper-secondary metacognitive utilisation schemes with GPT-4	Strong theoretical coding (instrumental genesis & MAEAB roles); pilot classroom interaction without post-withdrawal exam.
INC-48 Canonigo (2026)	JBI Qualitative	Secondary teacher critical mediation & equity in 2 Philippine schools	Strong interpretivist multi-school triangulation (10 teachers, 24 pupils, 8 weeks); qualitative perceptions/observations only.
INC-49 Oğuz Haçat & Kepceoğlu (2025)	JBI Qualitative	Grade 7 geographic/demographic math case study with ChatGPT (N = 4)	Exploratory small-sample case study (N = 4); documents verification vs. uncritical acceptance of AI calculations.

Note. Analytical cross-sectional studies were evaluated with the eight-item JBI manual checklist dated 26 March 2024.

S1.6 Reproducibility Files

Every pooled estimate runs from open code. Using All_36_Meta_Analytic_Contrasts.csv, standalone scripts in Python 3.11 (meta_stats_engine.py, run_final_meta_analysis.py) and R (reproduce_meta_analysis.R via metafor and clubSandwich) regenerate all cluster design-effect adjustments, Paule–Mandel and REML models, Hartung–Knapp–Sidik–Jonkman intervals, Egger regressions, Trim-and-Fill imputations, and Leave-One-Out refits.
Original audit notes, taxonomic codes, and bibliographic identifiers populate this supplement. Readers should consult the cited primary publications and open repositories for source texts; student-level microdata owned by third parties, copyrighted articles, and publisher abstracts are not reproduced here.