

Analysis of the use of artificial intelligence in mathematics education in Libya

Maryam Mohammed Alsoudani*

Department of Mathematics, Faculty of Education, University of Bani Waleed, Libya

*Email (for reference researcher): lymhmdalzwby838@gmail.com

تحليل استخدام الذكاء الاصطناعي في تعليم الرياضيات في ليبيا

مريم محمد السوداني اغنيه

قسم الرياضيات ، كلية التربية ، جامعة بني وليد ، ليبيا

Received: 02-08-2026; Accepted: 20-09-2026; Published: 25-09-2026

Abstract:

Artificial intelligence is entering mathematics education, but evidence about its use in Libya remains uneven. This study examines adoption, teaching applications, and learning claims across publicly available Libyan research. A targeted bilingual review identified eight relevant reports published between 2024 and September 2026. These included mathematics surveys, wider university surveys, documentary studies, and one reported teaching experiment. Published tables were extracted and checked through descriptive calculations. International experiments provided comparison evidence, without serving as estimates for Libya. The selected mathematics surveys covered Gharyan, Zawia, and Al-Marj. The Al-Marj study included 337 teachers, with primary teachers forming 74.2% of its sample. Reported classroom-use scores were lower than several workload and curriculum-barrier scores. A broader Zawia survey reported substantial training needs, alongside less frequent assessment use. However, most published percentages could not be reconstructed using its stated total sample alone. Another Zawia study contained a scale summary inconsistent with its displayed item means under ordinary averaging. A reported linear-equation experiment showed a nine-point difference in score changes between groups. Its reporting limitations prevented a firm causal interpretation. Overall, the evidence supports local engagement with AI, but cannot establish national adoption or sustained learning benefits. The paper proposes teacher-reviewed hints, independent mathematics assessment, and transparent reporting as priorities. Future Libyan studies should measure actual use, delayed learning, access differences, and teaching costs separately.

Keywords : artificial intelligence; mathematics education; Libya; teacher practices; evidence review; learning assessment..

المخلص:

يدخل الذكاء الاصطناعي تدريجياً إلى مجال تعليم الرياضيات، إلا أن الأدلة المتعلقة باستخدامه في ليبيا ما تزال غير متجانسة ومحدودة. تبحث هذه الدراسة في مدى تبني الذكاء الاصطناعي، وتطبيقاته في التدريس، والادعاءات المتعلقة بتأثيره على التعلم، وذلك من خلال مراجعة البحوث الليبية المتاحة للنشر العام. اعتمدت الدراسة مراجعة ثنائية اللغة موجهة، أسفرت عن تحديد ثمانية تقارير ذات صلة نُشرت خلال الفترة من عام 2024 حتى سبتمبر 2026. وشملت هذه التقارير دراسات مسحية في مجال الرياضيات، ومسوحاً جامعية أوسع نطاقاً، ودراسات وثائقية، وتجربة تعليمية واحدة تم الإبلاغ عنها. وقد جرى استخراج الجداول المنشورة والتحقق منها باستخدام حسابات وصفية. استُخدمت التجارب الدولية كأدلة للمقارنة فقط، دون اعتبارها تقديرات تمثل واقع استخدام الذكاء الاصطناعي في ليبيا. وشملت الدراسات المسحية المختارة في مجال الرياضيات مناطق غريان، والزواية، والمرج. وضمت دراسة المرج 337 معلماً، حيث شكّل معلمو المرحلة الابتدائية 74.2% من العينة. أظهرت درجات الاستخدام المبلغ عنها داخل الفصول الدراسية انخفاضاً مقارنة بدرجة مرتبطة بأعباء العمل ومعوقات المنهج الدراسي. كما أظهرت دراسة مسحية أوسع في الزاوية وجود احتياجات كبيرة للتدريب، إلى جانب انخفاض معدل استخدام الذكاء الاصطناعي في عمليات التقويم.

ومع ذلك، فإن معظم النسب المنشورة في هذه الدراسات لم يكن بالإمكان إعادة حسابها اعتمادًا على إجمالي حجم العينة المذكور فقط. كما تضمنت دراسة أخرى في الزاوية ملخصًا لمقياس قياس لا يتوافق مع متوسطات الفقرات المعروضة عند تطبيق أسلوب المتوسط الحسابي المعتاد. وأظهرت تجربة تعليمية مبلّغ عنها حول حل المعادلات الخطية فرقًا مقداره تسع نقاط في تغير الدرجات بين المجموعتين. إلا أن محدودية عرض البيانات في الدراسة حالت دون الوصول إلى تفسير سببي مؤكد للنتائج. وبصورة عامة، تدعم الأدلة المتاحة وجود اهتمام محلي بالذكاء الاصطناعي في التعليم، لكنها لا تكفي لإثبات مستوى التبني الوطني أو وجود فوائد تعليمية مستدامة. وتقتصر الورقة إعطاء الأولوية لاستخدام التلميحات التعليمية التي يراجعها المعلمون، وإجراء تقييمات مستقلة لمهارات الرياضيات، وتعزيز الشفافية في عرض البيانات. وينبغي للدراسات الليبية المستقبلية أن تقيس بصورة منفصلة: الاستخدام الفعلي للذكاء الاصطناعي، وتأثيره على التعلم طويل المدى، والفروق في إمكانية الوصول إليه، والتكاليف المرتبطة باستخدامه في العملية التعليمية.

الكلمات المفتاحية: الذكاء الاصطناعي؛ تعليم الرياضيات؛ ليبيا؛ ممارسات المعلمين؛ مراجعة الأدلة؛ تقويم التعلم.

Introduction

AI tools can produce worked solutions, explain symbols, and respond to mathematical questions. Their educational value depends on their use. A correct answer may support understanding, yet copying that answer may leave understanding unchanged. Mathematics education therefore requires evidence about reasoning, independent work, and transfer to unfamiliar problems.

Libya offers a useful setting for examining these questions. Recent local publications discuss AI within schools and universities. Their findings deserve attention, but their designs support different conclusions. This paper separates reported use, perceived benefits, and measured performance throughout its analysis.

The research problem

The central problem concerns what existing evidence can establish about AI in Libyan mathematics education. An adoption survey answers a different question from a teaching experiment. Agreement with a questionnaire statement cannot establish that students learned more mathematics. Likewise, access to an AI application does not show its regular use during lessons.

This distinction matters for institutional decisions. A university might purchase software because lecturers expect better teaching. That expectation could justify a pilot, but cannot demonstrate learning benefits. A school might report frequent use without recording the quality of feedback. These measures describe different parts of educational practice.

The selected local literature includes several kinds of evidence. Saleh (2025) studies mathematics teachers in Al-Marj. Salem and Milad (2025) examine mathematics teaching within Gharyan University. Mohamed (2025) considers faculty views at Zawia. Wider university surveys provide context, while documentary papers discuss possible uses and constraints. Their combined value lies in comparison, rather than numerical pooling.

Purpose and research questions

This paper examines the extent and forms of documented AI use in Libyan mathematics education. It also checks whether published data support their accompanying claims. Three questions guide the analysis. What forms of AI use appear in the selected local evidence? How strongly do those studies support adoption and learning claims? What teaching arrangements and research designs follow from the available evidence?

The contribution is a critical synthesis with reproducible calculations. The paper brings Arabic and English sources into one evidence map. It also examines selected tables, rather than accepting abstract conclusions alone. Original figures show extracted values and clearly marked secondary calculations.

The study does not estimate a national prevalence rate. Its sources lack a common sampling frame, measurement instrument, and observation period. Combining their percentages would therefore produce an unclear quantity. The paper instead identifies supported findings, unresolved reporting issues, and practical research priorities.

Boundaries of the analysis

The focus includes school mathematics and university mathematics education. Wider higher-education studies enter only as contextual evidence. Researcher affiliation does not establish the location of a field experiment. A mathematics department affiliation also does not make a general technology study mathematics-specific.

The review concerns published evidence available by September 28, 2026. Publication years do not necessarily identify data-collection years. Software capabilities also change between releases. Consequently, every finding belongs to its reported setting, period, and measurement approach.

Conceptual approach

What counts as AI use?

AI in this paper means software performing tasks through learned models, automated inference, or adaptive instructional decisions. Generative chatbots and intelligent tutoring systems fit this definition. Ordinary file sharing, slides, and basic calculators do not automatically qualify. A product containing AI should be classified through its specific instructional function.

The distinction prevents inflated adoption claims. Uploading a worksheet and receiving adaptive feedback involve different processes. Both use digital technology, but only the second necessarily involves an adaptive instructional decision. Researchers should record the product, feature, model version, task, and date of use.

Kulik and Fletcher (2016) reviewed 50 controlled evaluations of intelligent tutoring systems. Their work concerns structured tutoring systems, rather than current general-purpose chatbots. The distinction matters because instructional design forms part of the treatment. Evidence about one system cannot establish the effect of another.

Separating use from learning

This review applies four measurement categories. Awareness concerns knowledge of AI tools. Adoption concerns actual use within a stated period. Teaching practice concerns what teachers or students do with those tools. Learning concerns changes in mathematical knowledge or skill.

Table 1 Operational meanings used in this review.

Category	Suitable evidence	Insufficient evidence alone
Awareness	Identification or knowledge questions	Product access
Adoption	Dated use reports or activity records	Willingness to use
Teaching practice	Tasks, lesson observations, or reviewed logs	General approval
Assisted performance	Scores obtained while AI remains available	Claims about independent learning
Independent learning	Unaided tests, transfer, and delayed assessment	Faster completion with assistance

These categories may relate, but they remain distinct. A confident teacher may rarely use AI. A frequent user may ask only for final answers. An assisted score may improve without a corresponding increase in independent skill. Research instruments should therefore avoid placing all these meanings inside one overall score.

Why mathematical reasoning needs separate assessment

Mathematical work involves assumptions, transformations, and checks. A fluent explanation may hide an invalid transformation. Students need opportunities to identify such errors and justify alternatives. Teachers also need evidence that students understand why a method works. A useful assessment therefore includes several response types. These might involve selecting a method, explaining a step, correcting an error, and solving a new problem. Such tasks should remain aligned with the taught curriculum. They should also distinguish reading difficulties from mathematical difficulties where possible.

Methods

Review design and search

The study uses a targeted critical review with secondary descriptive analysis. It is not a preregistered systematic review. Searches combined English and Arabic topic terms with Libya

and institutional names. Publisher pages, university sites, proceedings, and accessible full texts were examined. International studies were selected for experimental relevance and accessible methods.

Searches included “artificial intelligence mathematics education Libya” and “الذكاء الاصطناعي الرياضيات ليبيا”. Follow-up terms included Gharyan, Zawia, Al-Marj, Tripoli, and linear equations. Exact-title searches located original publications after discovery.

Eligible local reports addressed mathematics education in Libya or supplied closely related university context. A Libyan-affiliated mathematics experiment was retained separately because its recruitment setting was unclear. Purely technical equation-solving research was excluded without educational participants or teaching analysis. Advertising and unsourced claims were excluded.

Table 2 Search and selection boundaries.

Component	Applied boundary
Languages	Arabic and English
Local publication period	2024 to September 28, 2026
Discovery	Web searches and institutional publication pages
Verification	Publisher records and accessible primary texts
Local synthesis	Eight selected reports; not an exhaustive census
International comparison	Two experiments and an established tutoring review
Exclusion example	Retracted ChatGPT meta-analysis; no effects extracted

Publication status formed part of checking. Wang and Fan (2026) report the retraction of their earlier ChatGPT meta-analysis. Its effect estimates were excluded. This check prevents a widely circulated result from entering the synthesis after withdrawal.

Extraction and synthesis

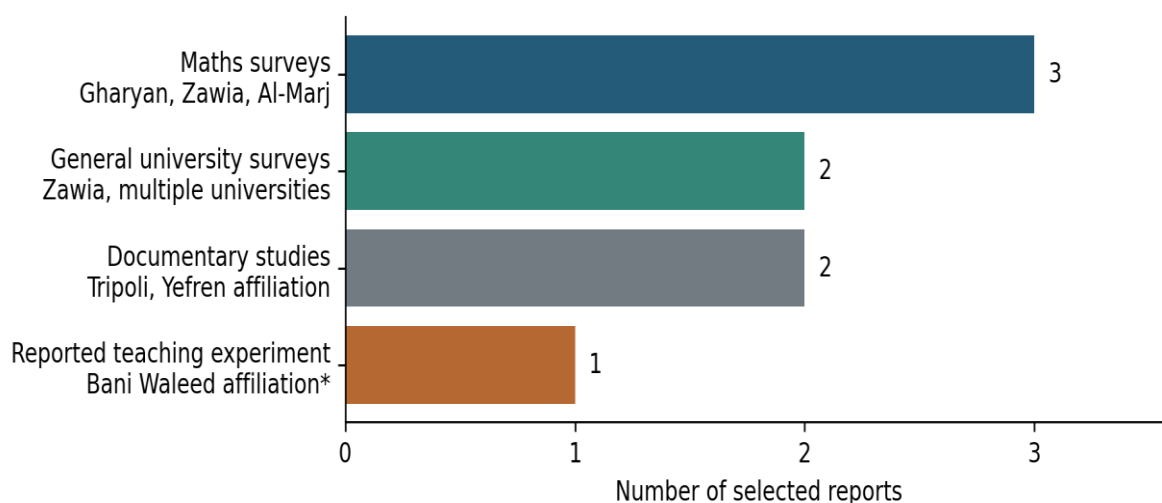
Extraction recorded setting, participants, design, sample size, measures, results, and reporting limitations. Arabic table labels were translated into short English descriptions. Visually checked tables took priority when extracted PDF numerals were distorted. The figures reproduce selected values, rather than entire source graphics.

The analysis uses descriptive comparisons and arithmetic checks. It does not pool effects across studies. Campbell et al. (2020) support clear reporting when quantitative results cannot be combined through meta-analysis. That reporting principle informs this paper, without implying full SWiM compliance.

Selection and extraction lacked independent duplicate review. Consequently, the paper remains vulnerable to missed studies and interpretation errors. The tables and links expose the evidence for further checking. No participant-level data were available for the local calculations.

Evidence profile

Eight local or locally affiliated reports form the selected evidence base. Their designs differ substantially. Figure 1 classifies them by their main contribution. The categories describe this review's sources, not the distribution of research across Libya.



*Recruitment site not established from the report.

Figure 1 Design profile of the eight selected reports.

Table 3 Selected local evidence and its scope.

Study	Setting or affiliation	Evidence and reported sample
Salem and Milad (2025)	Gharyan University	Mathematics survey; 27 respondents
Mohamed (2025)	University of Zawia	Mathematics faculty survey; 80 respondents
Saleh (2025)	Al-Marj public schools	Mathematics teacher survey; 337 respondents
Baroud et al. (2024)	University of Zawia	General university survey and interviews; 119 respondents
Abolkasim and Hasan (2024)	Libyan universities	General ChatGPT survey; over 1,000 responses
Al-Rakhis and Al-Hussan (2024)	University of Tripoli affiliation	Documentary mathematics analysis
Albaden (2026)	Yefren institute affiliation	Documentary mathematics analysis
Abu Naejah (2026)	Bani Waleed University affiliation	Reported experiment; 60 students; field site unclear

Only three reports directly survey local mathematics populations. Two broader surveys support contextual interpretation. Two documentary studies offer arguments without new participant observations. The remaining paper reports a teaching experiment, but its location requires clarification.

These distinctions prevent double counting of evidence types. A review's description of an earlier experiment does not create another local experiment. Similarly, participant totals cannot be added into one national sample. Recruitment methods differ, and overlap between institutional respondents cannot be ruled out.

School mathematics evidence from Al-Marj

Saleh (2025) reports 337 mathematics teachers from three school stages. Figure 2 shows the sample distribution. The study describes random selection, but national inference remains inappropriate. Its sampling location is one city, and the sample contains many primary teachers.

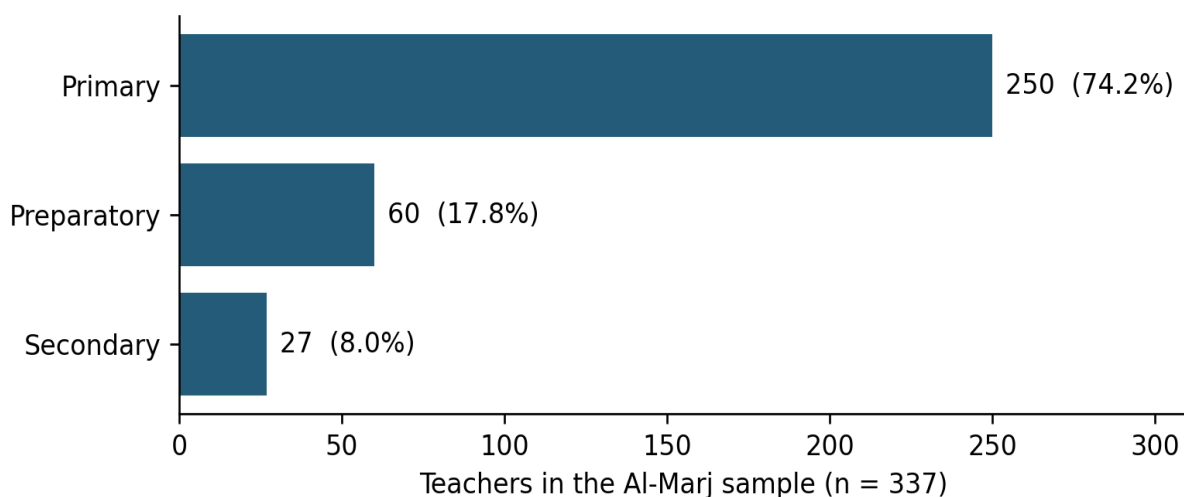


Figure 2 School-stage composition of the Al-Marj sample. Counts: Saleh (2025), Table 1, p. 13. Percentages are recalculated.

Primary teachers account for 250 respondents, or 74.2%. Preparatory teachers account for 60, or 17.8%. Secondary teachers account for 27, or 8.0%. These shares describe respondents, rather than AI adoption.

The composition affects interpretation. Primary mathematics often involves different instructional tasks from advanced secondary mathematics. A combined mean can conceal those differences. The study's overall results should therefore be read alongside its school-stage composition.

The percentages follow a simple calculation. Each group count is divided by 337 and multiplied by 100. Rounded values sum to 100.0%. No sampling weights were applied in this secondary description.

What the sample can establish

A city-level teacher survey can identify reported practices and concerns within its sampled population. It cannot establish what students learned from those practices. Nor can a teacher's estimate substitute for a student assessment. These boundaries remain even when the questionnaire contains many respondents.

For future replication, the sampling frame should identify eligible schools and teachers. The report should show invitations, responses, exclusions, and missing answers. Researchers should also explain whether multiple teachers came from the same school. Shared school conditions can make their responses more similar.

The Al-Marj evidence is useful because it concerns mathematics teachers directly. Its value does not depend on treating it as nationally representative. It can guide locally relevant questions for a broader study. Those questions should retain separate measures of knowledge, use, and constraints.

Knowledge, classroom use, and reported barriers

Figure 3 displays selected Al-Marj item means. General AI knowledge scored 3.63, while using AI for mathematics content scored 3.08. Teacher workload scored 4.46 as a barrier. Curriculum density scored 4.19, and limited lesson time scored 4.01 (Saleh, 2025).

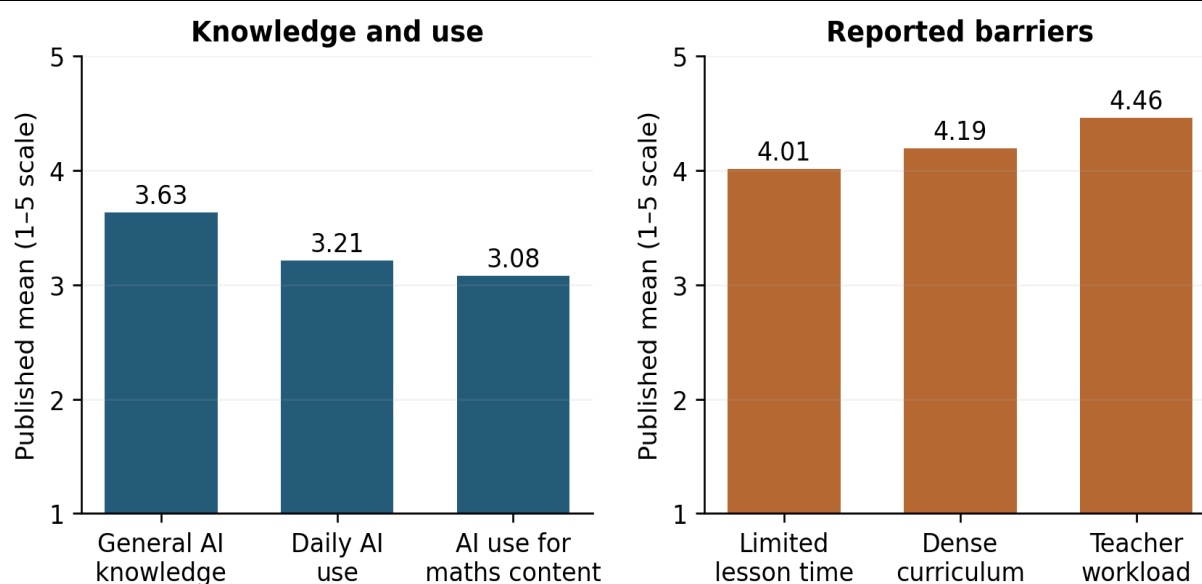


Figure 3 Selected AI-Marj means on a five-point scale. Source: Saleh (2025).

The descriptive pattern suggests a gap between familiarity and classroom application. It does not identify the cause of that gap. The selected items measure different constructs, so their differences are not treatment effects. The chart also cannot establish statistical significance between paired responses.

The original participant records would be needed for stronger comparisons. A paired analysis requires the relationship between each teacher's responses. Separate item means and standard deviations do not provide that relationship. This paper therefore reports the pattern without a new significance test.

Implications for teacher development

Training should connect AI use with actual mathematical lessons. General demonstrations may leave teachers unsure about lesson timing and curriculum fit. A practical session could begin with one common misconception. Teachers could compare several hints and explain which one addresses that misconception.

Workload should also be measured during a pilot. Preparing prompts, checking outputs, and managing devices all require time. Faster material generation may not reduce total preparation time. Any claimed saving should include checking and correction.

Curriculum fit can be assessed through a small lesson sequence. Teachers can identify where AI adds useful feedback and where it adds little. The resulting record should include abandoned activities and technical failures. Reporting only successful demonstrations would provide an incomplete account of classroom use.

University evidence: Gharyan and Zawia

Salem and Milad (2025) report 27 respondents from Gharyan University's education faculties. The sample includes 23 students and four faculty members. Thus, student responses dominate a study partly addressing faculty practices. Its questionnaire findings should not be described as direct observation of teaching frequency.

The small faculty subgroup limits claims about lecturers. Four respondents cannot describe institutional variation reliably. Combining staff and student responses also changes the meaning of an overall mean. Future reporting should display both groups separately, with their exact question wording.

Baroud et al. (2024) examine wider university practice at Zawia. Figure 4 shows selected published percentages. The stated sample contains 119 respondents across specialties. These results provide institutional context, without identifying mathematics-specific use.

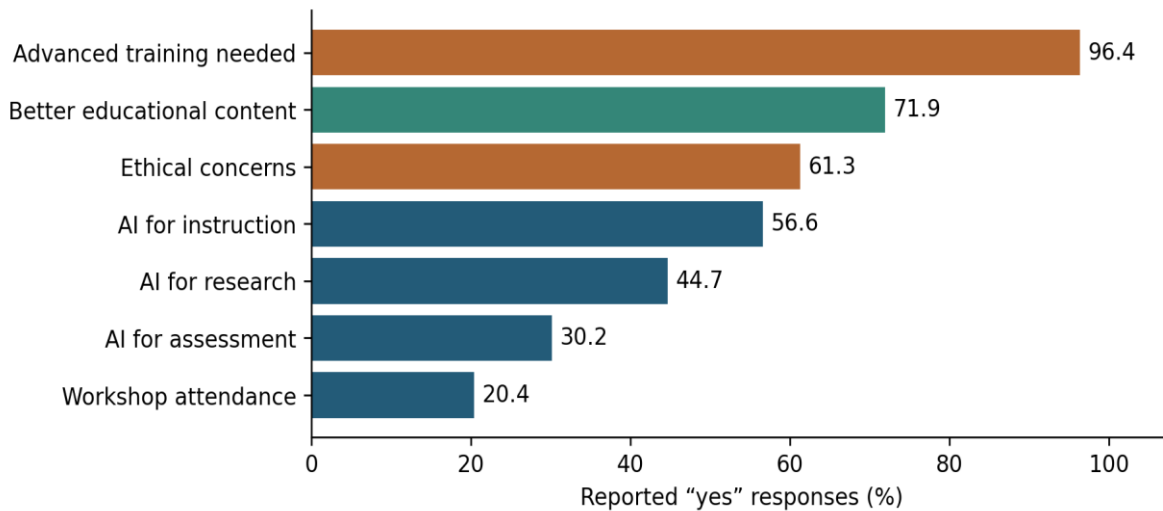


Figure 4 Selected reported responses at Zawia. Source: Baroud et al. (2024).

The percentages require a denominator check. Twelve of the table's 13 percentages cannot result from whole-person counts over 119. This finding assumes unweighted responses and ordinary rounding to one decimal place. Item-specific denominators or weighting could change that conclusion.

The discrepancy does not establish invalid data. It identifies missing information needed for reconstruction. Counts, missing responses, and weighting rules would resolve the issue. Until then, the published percentages should remain descriptive reports, without newly calculated confidence intervals.

Checking the Zawia mathematics scale

Mohamed (2025) surveyed 80 faculty members. Table 6 displays 11 means for a scale labelled mathematics-learning improvement. Its items mainly concern barriers or negative statements. The displayed means range from 2.90 to 3.48, while the reported overall mean is 3.88.

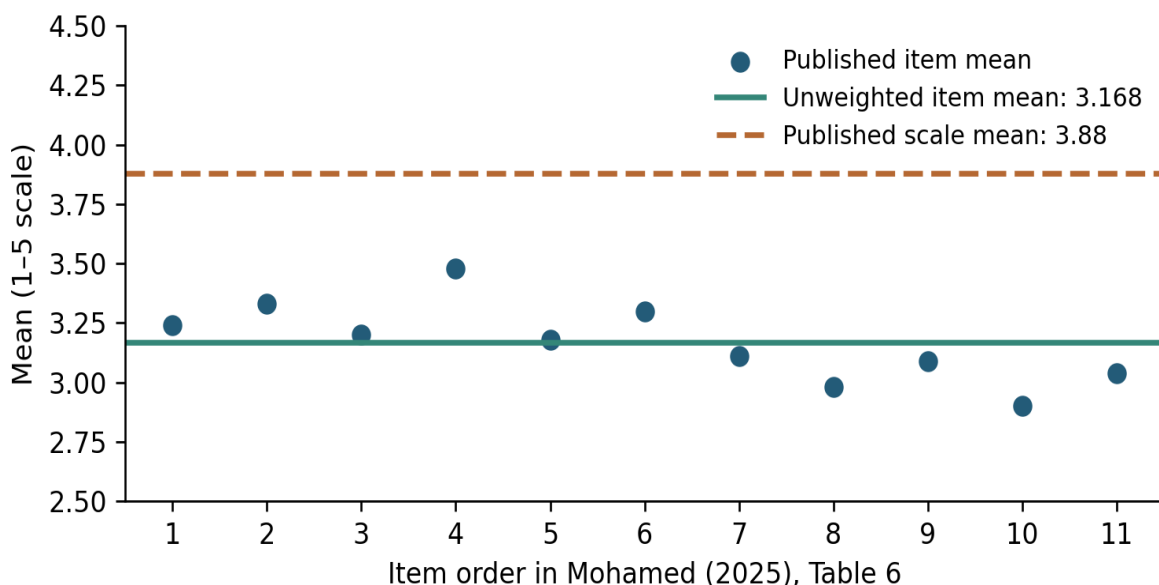


Figure 5 Audit of the displayed mathematics-learning scale. Source: Mohamed (2025).

The unweighted average of those displayed means is 3.168. The gap from 3.88 is approximately 0.712. Under common respondents and unchanged scoring, ordinary averaging cannot produce the reported overall value. Any nonnegative weighted average must also remain within the item range.

This check has explicit limits. Different missing-data rules, reverse scoring, or another transformation could affect a composite. The displayed summary does not provide a

reconciliation. Therefore, the calculation identifies an unresolved reporting issue, rather than a corrected participant-level scale.

The construct also requires attention. Agreement that training is needed does not measure better mathematical understanding. A barrier scale can be useful, but its label should match its content. Reverse coding alone cannot turn a barrier opinion into an achievement test.

Table 4 Selected checks and their interpretation.

Check	Secondary result	Required clarification
Zawia percentage reconstruction	12 of 13 values incompatible with a common denominator of 119	Item denominators and weights
Mathematics scale averaging	Displayed-item mean 3.168; published summary 3.88	Scoring and missing-data rules
Scale interpretation	Barrier items are not achievement scores	Construct definition and validation

The broader lesson concerns the link between measurement and claims. A statistically significant association between questionnaires remains an association between questionnaire scores. It does not establish that AI caused better learning. Direct student evidence is needed for that conclusion.

The reported linear-equation experiment

Abu Naejah (2026) reports two groups of 30 students and a 90-day intervention. Figure 6 redraws the published test summaries. The author's affiliation is Bani Waleed University. The report does not clearly establish the participant recruitment site.

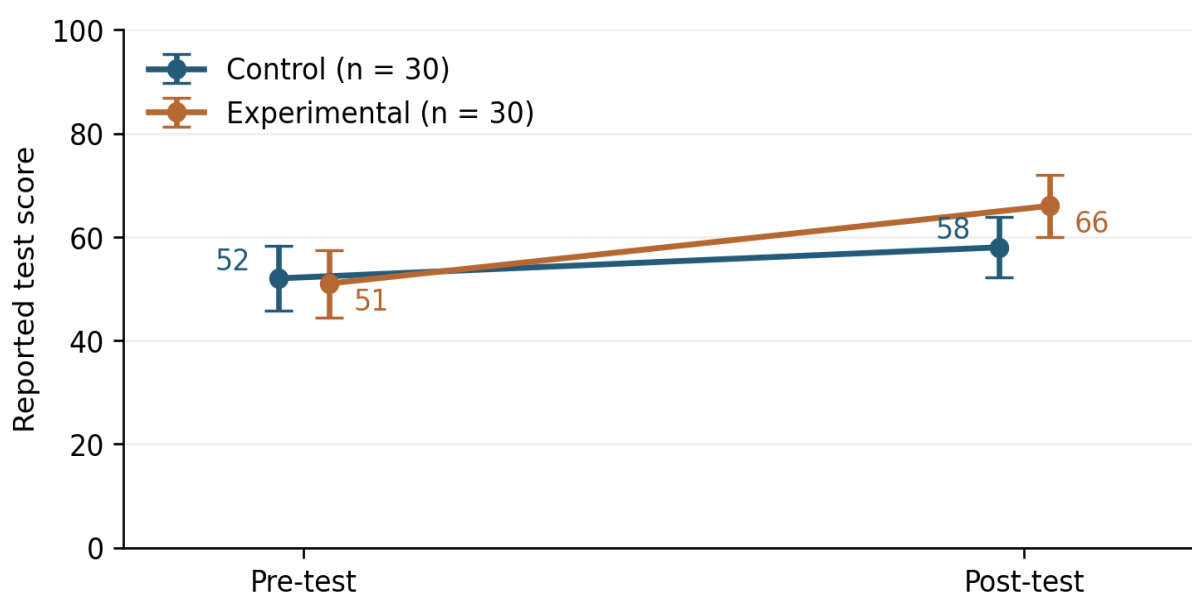


Figure 6 Published pre-test and post-test scores. Source: Abu Naejah (2026).

The control mean rises from 52 to 58. The experimental mean rises from 51 to 66. Their respective changes are six and 15 points. The difference between those changes is nine points. This descriptive contrast is not an adjusted causal estimate.

The report creates uncertainty about the treatment contrast. Its English abstract describes AI training for both groups. Its Arabic abstract describes training for the experimental group. The methods also indicate AI availability during experimental testing. Thus, assisted performance may contribute to the observed difference.

No new confidence interval is calculated for the change contrast. Such an interval requires information about the covariance between repeated scores. Group allocation, attrition, and assessment conditions also require clearer documentation. These limitations prevent a firm claim about independent learning.

How the experiment should inform practice

The reported means justify further investigation of equation teaching. They do not justify predicting a particular gain for another class. A replication should use identical assessment conditions for both groups. It should also separate practice with assistance from testing without assistance.

Participants should solve unfamiliar equations after the teaching period. They should explain transformations and check solutions through substitution. A later assessment should measure retention. Recording these outcomes would reveal whether any benefit continues when the tool disappears.

International experimental comparisons

Bastani et al. (2025) studied high-school mathematics in Turkey. Their experiment compared ordinary instruction with two GPT-based support arrangements. Figure 7 shows adjusted differences from the control condition. The assisted and unassisted assessments produce different patterns.

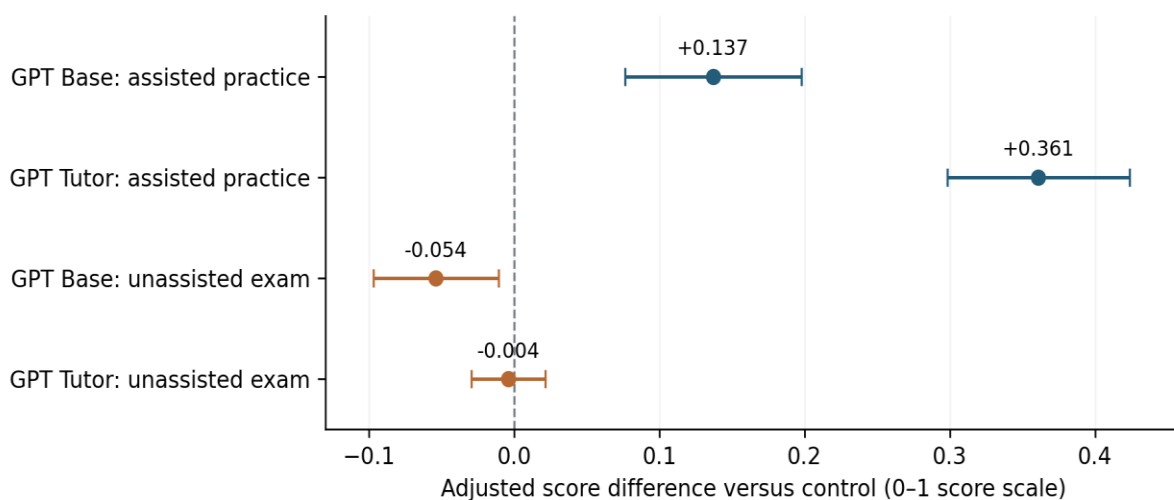


Figure 7 Adjusted effects in the Turkish experiment. Source: Bastani et al. (2025).

For assisted practice, the reported effects are 0.137 and 0.361 on a zero-to-one score scale. For unaided exams, they are -0.054 and -0.004 . The first exam effect concerns GPT Base. The second concerns GPT Tutor. The source uses classroom-clustered standard errors.

These results explain why assisted performance and independent learning require separate outcomes. They do not show what would happen in Libya. Language, curriculum, prior knowledge, and classroom conditions could alter results. The comparison instead informs measurement and instructional design.

Wang et al. (2025) study Tutor CoPilot, which supports human tutors. Their preregistered trial reports a four-percentage-point increase in topic mastery. Students of lower-rated tutors show a nine-point increase. The revised preprint describes 900 tutors and 1,800 students. These figures concern overlapping groups and should not be pooled.

The two experiments involve different interventions and outcomes. One provides student-facing support; the other assists tutors. Neither establishes that every chatbot improves mathematics learning. Their shared value is the attention given to instructional arrangements and measurable outcomes.

Public practical materials accompany both projects. Bastani and colleagues provide a data-and-code repository. Tutor CoPilot provides demonstration code and tutorials. Table 6 identifies those resources. Access to published materials supports replication planning, but does not mean this review reran either trial.

Discussion

What can be said about the extent of use?

The selected evidence supports local engagement with AI in mathematics education. It does not establish a national adoption rate. Several studies measure views, confidence, or perceived benefits alongside use. Their findings should be interpreted according to the actual questions asked.

A meaningful adoption estimate needs a clear denominator. For example, researchers might examine all mathematics teachers within a defined district. They would specify what counts as AI use during a stated teaching period. The resulting estimate would still describe that district and period.

Institutional adoption requires another measure. A lecturer's private chatbot account does not establish university integration. Integration could require approved access, training, curricular use, and technical support. Researchers should distinguish personal experimentation from supported institutional practice.

Use also varies by task. Preparing examples, marking work, giving hints, and producing complete solutions involve different educational risks. One overall usage score conceals these differences. A useful survey should therefore report task-specific frequencies and representative examples.

Why favourable opinions remain insufficient

Teachers may reasonably value immediate explanations and draft materials. Such preferences can help identify acceptable teaching arrangements. However, satisfaction does not establish mathematical accuracy or learning. Students may also prefer an activity because it reduces effort.

The distinction becomes clear when assessment conditions change. During assisted work, students and software jointly produce the answer. During an unaided test, the student must supply the reasoning. Comparing these conditions helps identify what students learned and what the tool supplied.

This principle applies beyond chatbots. An adaptive platform may improve practice completion without improving transfer. A diagnostic system may classify errors accurately without helping teachers address them. Educational evaluation must examine the complete teaching process, including the response to feedback.

Reading local documentary studies

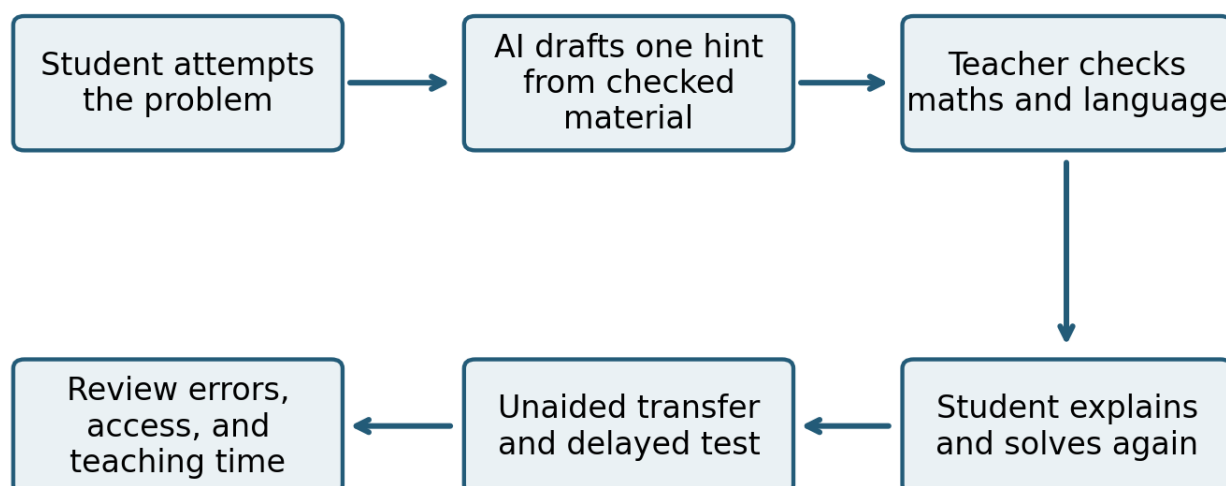
Al-Rakhis and Al-Hussan (2024) discuss mathematics opportunities, privacy, and difficulties assessing deep understanding. Albaden (2026) considers teaching strategies and implementation constraints. Both use documentary analysis, rather than local field experiments. Their arguments help frame questions but do not measure adoption.

Abolkasim and Hasan (2024) provide wider university context through a survey exceeding 1,000 responses. That broader scope does not supply a mathematics-specific denominator. Consequently, this review keeps their findings outside the mathematics prevalence analysis. The evidence therefore supports a measured institutional response. Small, documented pilots are easier to evaluate than broad claims about transformation. Their purpose should be to improve a defined mathematical skill. Continued use should depend on results, feasibility, and teacher judgment.

A practical teaching application

Teacher-reviewed hints for linear equations

Figure 8 presents a proposed classroom sequence. It is an instructional design, not an evaluated Libyan intervention. Its central rule is simple: students attempt the mathematics before receiving a hint. The teacher checks the hint before using it with the class.



Proposed classroom sequence; no trial results are implied.

Figure 8 Proposed teacher-reviewed hint sequence

A suitable example asks students to solve $3(x - 2) = 2x + 5$. The correct solution is $x = 11$. A teacher can verify this by substituting into both sides. Both expressions then equal 27.

Suppose a student expands the left side as $3x - 2$. The teacher can request a hint about the distributive property. The hint should identify the relevant idea without supplying the final answer. Students then explain why both terms inside the bracket require multiplication.

An example prompt is: "Give one short hint about the first incorrect step. Do not provide the final answer." This prompt is proposed wording, not a tested safeguard. A model might still reveal the solution or produce an incorrect explanation. Teacher review remains necessary.

Recording the learning process

The lesson record should retain the initial attempt, reviewed hint, revised solution, and student explanation. A short unaided task should follow. For example, students could solve $4(x - 3) = 2x + 8$ independently. The correct answer is $x = 10$.

The teacher should also record preparation and checking time. These records help assess whether the activity is manageable. Students without personal devices should receive an equivalent teacher-mediated version. This arrangement permits participation without assuming equal private access.

A future Libyan evaluation design

A stronger study could begin within a mathematics department or a defined group of schools. Institutional cooperation would be needed before recruitment. The following design is proposed for future work. No such recruitment or intervention occurred during this review.

The study would compare usual teaching with teacher-reviewed AI hints. Both groups would receive equal curriculum coverage and practice time. The primary outcome would be an unaided mathematics assessment. A delayed assessment would examine retention and transfer.

Table 5 Proposed measures for a future evaluation.

Dimension	Measurement	Interpretation
Actual use	Dated task logs and teacher records	Exposure and task type
Independent learning	Common unaided assessment	Main educational outcome
Transfer	Unfamiliar problems using taught concepts	Application beyond practice items
Retention	Delayed unaided assessment	Persistence of learning
Mathematical errors	Blind review using a shared rubric	Accuracy and reasoning
Access	Device, connection, and language records	Unequal participation
Workload	Preparation, checking, and lesson minutes	Practical teaching cost

Random allocation would strengthen causal interpretation where feasible. Class-level allocation could reduce exchange between conditions. However, analysis would then need to account for clustering. Sample size should follow the primary outcome and expected cluster structure. This paper does not invent a required participant count.

Where randomization is impossible, comparison groups should receive the same assessment schedule. Baseline scores and relevant background measures should be recorded. The analysis should adjust for planned differences without claiming that adjustment removes every bias. Attrition should be reported by condition.

The assessment rubric should reward explanation and valid transformations. It should also recognize correct alternative methods. Markers should assess anonymized work without knowing the teaching condition when possible. Agreement between markers should be reported.

The analysis plan should specify the main comparison before examining outcomes. Researchers should report estimates and uncertainty, including results that do not favour AI. Subgroup analyses should remain limited and justified. Multiple exploratory comparisons should not replace the main outcome.

Ethics approval and informed consent should precede participant work. Student records should remain separate from public AI prompts. For younger pupils, the institution should establish suitable consent and supervision arrangements. These steps concern the proposed study, rather than approvals claimed for this review.

Institutional and language considerations

Arabic mathematics and teacher checking

Arabic instruction creates specific design questions. Mathematical symbols may appear beside right-to-left text. A technically correct answer can still confuse students through layout or unfamiliar terminology. These possibilities should be tested with actual lesson materials.

Evaluation should include Arabic prompts, mixed notation, and locally used textbook language. Teachers can compare explanations for equivalent questions in different forms. The aim is to identify errors that matter for teaching. General language fluency alone is an insufficient measure.

Institutions should retain a reviewed set of examples for common topics. This set could include correct solutions and frequent misconceptions. It would support consistent checking across teachers. It should also contain problems where a tool previously failed.

Access and workload

An institution should measure access before choosing an instructional arrangement. The relevant questions concern dependable classroom access and support. Ownership of a phone alone does not establish either condition. Researchers should record interrupted sessions and unavailable features.

A teacher-mediated activity may reduce dependence on individual accounts. It also permits collective discussion of a shared explanation. However, that arrangement changes the intervention. Results from individual tutoring cannot be assumed to apply unchanged.

Cost evaluation should include training, technical support, subscriptions, and checking time. It should also include the cost of alternatives. A free application may still create substantial staff work. Decisions require total teaching costs, rather than subscription prices alone.

Privacy and academic judgment

Miao and Holmes (2023) emphasize human oversight, privacy, and appropriate educational use of generative AI. Those principles support clear institutional rules. Teachers should know which data may enter a service and who remains responsible for decisions.

For mathematics assessment, automated suggestions should remain reviewable. A student should be able to challenge an incorrect explanation or mark. Institutions should preserve the original work and the teacher's final decision. This record supports both fairness and later research.

Academic integrity rules should describe acceptable assistance in each task. Students need to know whether hints, checking, or generated solutions are permitted. Requirements should match the learning purpose. An unaided test serves a different purpose from collaborative practice.

Policy should also address changing software. Institutions should record versions and dates during an evaluation. A later update may change mathematical accuracy or response style. That change creates a new evaluation question, rather than automatically confirming earlier findings.

Limitations

This review is targeted and cannot claim complete coverage. Web discovery may favour accessible and well-indexed publications. Arabic search variants and inconsistent metadata may hide relevant studies. Some local sources also report limited methodological detail.

The extraction was not independently duplicated. Translated labels may lose some nuance, despite checking the displayed tables. Public summaries cannot reveal participant-level distributions or missing-data patterns. Accordingly, this paper avoids new causal estimates from local survey results.

The arithmetic checks depend on stated assumptions. Percentage reconstruction assumes a common denominator and unweighted responses. The scale check assumes comparable item scoring and respondents. Different procedures could explain a discrepancy, but require explicit documentation.

The selected experiment has unresolved reporting and setting questions. Its inclusion does not certify its findings. International experiments offer stronger design examples, but do not establish Libyan effects. School and university populations also require separate interpretation. The eight figures serve different purposes. Some display reported observations, while others display calculations or review coding. The final diagram presents a proposal. None depicts a new Libyan experiment performed for this manuscript.

Conclusion

Published evidence shows interest and reported AI use within several Libyan educational settings. Mathematics-specific evidence remains concentrated in surveys and a small number of reports. It cannot support a national adoption percentage or a firm claim of sustained learning gains.

The most useful next step is careful measurement within actual mathematics lessons. Studies should distinguish awareness, task-specific use, assisted performance, and independent learning. Reports should provide exact denominators, scoring rules, assessment conditions, and uncertainty.

Teacher-reviewed hints offer one practical arrangement for testing. They connect AI assistance with mathematical explanation and subsequent unaided work. Their value in Libya must still be evaluated. Institutional decisions should follow those results, alongside access, workload, and student needs.

References

1. Abu Naejah, K. N. A. (2026). The role of artificial intelligence in enhancing linear equation modeling and its applications for science-stream students. *Scientific Journal for Publishing in Health Research and Technology*, 2(2), 77–92. <https://doi.org/10.65420/sjphrt.v2i2.122>
2. Abolkasim, E., & Hasan, M. (2024). Integrating ChatGPT in education and learning: A case study on Libyan universities. *Journal of Pure & Applied Sciences*, 23(2), 19–24. <https://doi.org/10.51984/jopas.v23i2.3082>
3. Albaden, F. S. M. (2026). Strategies for employing artificial intelligence (AI) technologies in the development of mathematics teaching: Challenges and future prospects. *International Science and Technology Journal*, 38. <https://www.stcrs.com.ly/istj/en/article-details.php?id=1416>

4. Al-Rakhis, T. A., & Al-Hussan, F. A. (2024). Al-takamul bayna al-dhaka al-istinai wa-tadris al-riyadiyah: Al-tahaddiyat wa-al-furas [Integration of artificial intelligence and mathematics teaching: Challenges and opportunities]. *Journal of the Faculty of Education Tripoli*, 1(20). <https://doi.org/10.5281/zenodo.17412571>
5. Ali, E. A. R., Naser, N. R., & Atlouba, N. A. (2025). Applications of matrices in big data analysis: Modeling user behavior in digital systems. *Libyan Open University Journal of Applied Sciences (LOUJAS)*, 41-52.
6. Algamaty, R. K. (2025). Fair and Transparent AI for Hiring: Evaluating Resume-Job Matching, Bias Mitigation, and Human-in-the-Loop Auditing. *Libyan Open University Journal of Applied Sciences (LOUJAS)*, 01-10.
7. Baroud, N., Alouzi, K., Elfzzani, Z., Ayad, N., & Albshkar, H. (2024). Educators' perspectives on using (AI) as a content creation tool in Libyan higher education: A case study of the University of Zawia. *JERIT: Journal of Educational Research and Innovation Technology*, 1(2), 61–70. <https://doi.org/10.34125/jerit.v1i2.12>
8. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
9. Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, 368, l6890. <https://doi.org/10.1136/bmj.l6890>
10. Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78. <https://doi.org/10.3102/0034654315581420>
11. Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
12. Mohamed, N. A. (2025). The use of artificial intelligence applications in enhancing mathematics learning from the perspective of faculty members at University of Zawia. *International Science and Technology Journal*, 36(1), 1–20. <https://doi.org/10.62341/NAMT2054>
13. Saleh, M. A. (2025). The potential of employing artificial intelligence applications in teaching mathematics and its challenges in public schools in Al-Marj City. In *Proceedings of the first international scientific conference of the Faculty of Education, Gharyan University (Vol. 2, pp. 1–28)*. Gharyan University. <https://edco.gu.edu.ly/ar/j/issue-articles/678/download>
14. Salem, R. A. M., & Milad, M. A. A. (2025). Darajat istikhdam taqniyat al-dhaka al-istinai fi tadris maddat al-riyadiyah bi-kulliyat al-tarbiya Jamiat Gharyan [The degree of using artificial intelligence technologies in teaching mathematics at the education faculties of Gharyan University]. *Al-Riyadah Journal for Researches and Scientific Activities*, 4(11), 60–80. <https://alriyadah.ly/uploads/topics/17581786089136.pdf>
15. Wang, J., & Fan, W. (2026). Retraction note: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, 13, Article 528. <https://doi.org/10.1057/s41599-026-07310-z>
16. Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2025). Tutor CoPilot: A human-AI approach for scaling real-time expertise (Version 2) [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2410.03017>

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of LOUJAS and/or the editor(s). LOUJAS and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.