

Developing and Evaluating an AI Prompt Blueprint for Interactive Seventh-Grade Mathematics Media

Habib Burrahman^{1,*}

¹Universitas Islam Negeri Sjech M. Djamil Djambek Bukittinggi, Bukittinggi, Indonesia

Article Information

Article History:

Submitted: March 30, 2026

Revised: June 10, 2026

Accepted: June 29, 2026

Published: June 30, 2026

Keywords

prompt engineering; Canva Code; Grade 7 mathematics; generative AI; ADDIE

Correspondence

E-mail:

habibburr.orabi@gmail.com*

A B S T R A C T

This study developed and evaluated a structured AI prompt blueprint for generating interactive Grade 7 mathematics learning media using Canva Code. Guided by the ADDIE model, the research and development project produced 14 prompts addressing lines and angles, plane geometry, and statistics. Five experts evaluated the prompts and generated media, three mathematics teachers completed a 22-item response questionnaire, and 27 students answered six statements before and after using the media. Normalized expert-rating summaries were 0.90 for the prompts and 0.91 for the media. The teachers' mean response was 4.61/5 (92.12% of the maximum possible score), while students' mean item responses increased from 1.98 to 4.20 (normalized score change = 0.73). The available aggregated records do not permit verification of item-level Aiken's V or calculation of Cohen's kappa, and the six student statements have not been established as a mathematics achievement test. Accordingly, the findings indicate favorable initial ratings and questionnaire responses but do not establish causal improvements in mathematical proficiency. The study contributes a documented prompt-design workflow that distinguishes prompt evaluation from evaluation of AI-generated instructional media and underscores the need for human review and further validation.

Abstrak

Penelitian ini mengembangkan dan mengevaluasi blueprint prompt AI untuk pembuatan media matematika interaktif kelas VII menggunakan Canva Code. Penelitian pengembangan dengan model ADDIE menghasilkan 14 rancangan prompt pada materi garis dan sudut, bangun datar, serta statistika. Lima ahli menilai rancangan prompt dan media; tiga guru mengisi 22 butir angket respons; dan 27 siswa mengisi enam pernyataan sebelum dan sesudah penggunaan media. Ringkasan penilaian ahli yang dinormalisasi adalah 0,90 untuk prompt dan 0,91 untuk media. Rerata respons guru adalah 4,61 dari 5 (92,12% skor maksimum). Rerata skor per butir siswa berubah dari 1,98 menjadi 4,20, dengan perubahan ternormalisasi 0,73. Data yang tersedia tidak memungkinkan verifikasi Aiken's V per butir maupun Cohen's Kappa; enam butir siswa juga belum dapat dipastikan sebagai tes prestasi matematika. Dengan demikian, temuan mendukung kelayakan awal rancangan dan respons positif dalam uji coba terbatas, tetapi belum membuktikan peningkatan hasil belajar secara kausal. Penelitian ini memberikan kontribusi berupa alur kerja perancangan prompt yang terdokumentasi – yang membedakan antara evaluasi prompt dengan evaluasi media pembelajaran hasil AI – serta menekankan pentingnya tinjauan oleh manusia dan validasi lebih lanjut.

This is an open-access article under the CC-BY-SA license.



1. Introduction

Generative artificial intelligence (GenAI) is reshaping how teachers prepare instructional materials, but the availability of generated content does not, in itself, establish its pedagogical suitability. Reviews of AI in education identify opportunities for content development, assessment, and feedback, alongside concerns about accuracy, equity, and classroom governance [1]–[4]. These concerns are particularly relevant in resource-constrained schools, where teachers may be able to generate a functional digital interface without having a reliable procedure for checking mathematical content or alignment with instructional objectives. A human-centered approach therefore requires explicit learning objectives, transparent instructions, and teacher oversight [5]–[7].

Grade 7 mathematics includes topics such as lines and angles, plane geometry, and statistics, in which visual representations and immediate feedback may support learning. The cognitive theory of multimedia learning explains how complementary verbal and visual information can promote active cognitive processing when instructional materials are appropriately designed [8], [9]. Previous studies of Canva-based mathematics media have reported promising classroom applications [10]–[12]. Research on digital mathematics agents and AI-generated hints also suggests potential benefits while highlighting risks related to content quality [13]–[16]. These findings do not imply that every automatically generated resource is mathematically accurate or pedagogically effective.

Prompt engineering is increasingly understood as part of instructional design rather than merely the technical task of writing commands. Systematic reviews emphasize structured prompting, iterative refinement, and alignment with learning objectives [17]–[20]. Empirical studies of AI-generated lesson plans likewise underscore the importance of explicit pedagogical constraints and expert appraisal [21]–[25]. Research in middle-school mathematics has documented limitations in the responses of general-purpose AI systems to challenging tasks, further emphasizing the need for subject-specific verification [26], [27]. However, the literature reviewed does not demonstrate that the particular Role–Context–Task–Constrain–Output structure used here reduces the frequency of AI hallucinations.

This study addresses a practical and methodological question: how can a structured prompt blueprint be documented and evaluated as a development artifact alongside the instructional media it generates? The contribution is limited to the design and preliminary evaluation of 14 prompts for mathematics media at SMP Negeri 3 Ampek Angkek, Indonesia. Specifically, the study examines (1) how the prompts were structured and refined; (2) how experts evaluated the prompts and generated media; (3) how teachers responded to the media; and (4) how students' six-item questionnaire scores differed before and after use. Unlike research testing achievement effects through controlled designs [13], [16], this single-group trial provides only short-term descriptive evidence. Distinguishing the prompt, generated product, and learner response is central to the study's contribution [28], [29].

2. Methodology

2.1. Research Design

This research adopted a research and development (R&D) approach structured around the five ADDIE phases: analysis, design, development, implementation, and evaluation [30]. This approach was appropriate for producing and refining digital instructional artifacts. An exploratory, single-group questionnaire assessment administered before and after media use was incorporated into the development process; it was not designed to establish causal effects. The four evaluation targets were prompt quality, the quality of generated media, teacher responses, and changes in students' questionnaire scores. Accordingly, validity, practicality, and effectiveness are treated as distinct constructs rather than interchangeable outcomes.

2.2. Setting and Participants

The study was conducted from May to July 2026 at SMP Negeri 3 Ampek Angkek, Agam Regency, West Sumatra, Indonesia. Participants comprised five expert assessors, three mathematics teachers, and 27 Grade 7 students. Student selection was described as random sampling, but the sampling frame and selection procedure were not documented. Consequently, the representativeness of this sample beyond the study group cannot be established.

- 2.2.1. Expert assessors: three media specialists with postgraduate qualifications in computing or computer education, one mathematics-content specialist experienced in teaching Grade 7, and one Indonesian-language specialist.
- 2.2.2. Teacher participants: three mathematics teachers who examined the generated instructional media and completed a questionnaire about their experiences using it.
- 2.2.3. Student participants: 27 Grade 7 students who had previously studied at least one of the relevant mathematics topics. Additional demographic characteristics, recruitment procedures, and participant allocation details were not reported.

2.3. Development Procedure

The reported development process followed the five ADDIE phases:

- 2.3.1. Analyze. The development team considered instructional media needs, characteristics of teachers and students, device and internet access, availability of Canva Education accounts, and institutional readiness.
- 2.3.2. Design. The team developed the RCTCO blueprint (Role, Context, Task, Constrain, Output), specified interactive content, and prepared assessment instruments. RCTCO served as the operational prompt structure for this project rather than as an independently established theoretical framework.
- 2.3.3. Develop. Prompts were entered into Canva Code and refined through iterative generation and review. The study reports between zero and five revisions per media item, primarily to address interface behavior, mathematical representation, and functionality. A complete item-by-item development log was not provided.
- 2.3.4. Implement. Five experts assessed the prompts and generated media, three teachers completed questionnaires about their experiences using the media, and 27 students completed the same six-item questionnaire before and after media use.
- 2.3.5. Evaluate. Expert judgments and participant responses were summarized as part of the final development evaluation. Neither a second independent evaluation cohort nor a delayed retention test was reported.

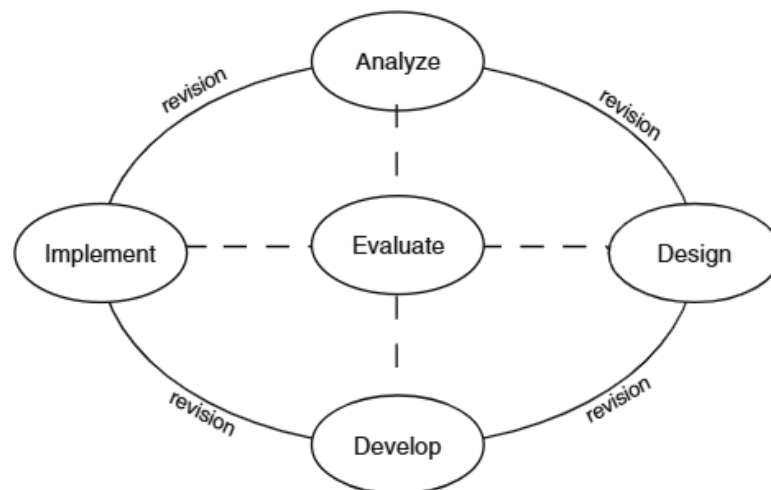


Figure 1. ADDIE development model (adapted from [30]).

2.4. Instruments

- 2.4.1. Expert-rating forms used five-point response scales (1–5) to evaluate visual and operational features, perceived usefulness, mathematical content, and language. The number of items varied according to the assessor's area of expertise and the artifact being evaluated. Item-level ratings were not reported.
- 2.4.2. Three mathematics teachers completed a 22-item response questionnaire concerning ease of use, prompt-design quality, and perceived usefulness. The available results provide only each teacher's total score, without item-level responses or the cross-tabulated categorical ratings required to estimate inter-rater agreement.
- 2.4.3. Students completed the same six-statement questionnaire before and after using the media. Neither the full item wording nor evidence that the instrument measured mathematics achievement was reported. It is therefore treated as a self-report questionnaire rather than as a validated mathematics achievement test.

2.5. Data Collection Procedures

Data collection proceeded in three stages: experts assessed the prompts and media; teachers used the media and completed the response questionnaire; and students completed the questionnaire before and after media exposure. The duration of classroom use, the materials assigned to individual students, the interval between assessments, the complete assessment rubrics, and procedures for handling missing item-level data were not specified. These details would need to be recovered from the original research records rather than reconstructed retrospectively.

2.6. Data Analysis

- 2.6.1. Expert assessments. Aiken's V measures item-level content validity using deviations from the lowest rating category, the number of judges, and the scale width [31]. The available tables report only assessor-level totals for instruments with unequal item counts. A normalized mean can be derived from these totals as $[\Sigma(\text{mean assessor score} - 1)]/[5 \times (5 - 1)]$, yielding 0.9025 for the prompt blueprints and 0.9111 for the generated media. The previously reported values of 0.90 and 0.91 are therefore retained as rounded, normalized descriptive ratings. Without the original item-level rating matrices, they cannot be independently verified as conventional Aiken's V coefficients.

$$V = \frac{\Sigma S}{[n(c - 1)]} \quad (1)$$

- 2.6.2. Teacher responses. Cohen's kappa measures agreement between categorical ratings after accounting for chance agreement [32]. The necessary categorical data were not available for the three 22-item teacher questionnaires. Thus, the originally reported "kappa = 0.91" cannot be verified as an inter-rater agreement statistic or treated as a valid measure of practicality. The reported questionnaire totals of 101, 99, and 104 (each out of 110) instead yield a descriptive pooled mean of $304/(3 \times 22) = 4.61/5$, equivalent to 92.12% of the maximum possible score.

$$K = \frac{P0 - Pe}{1 - Pe} \quad (2)$$

2.6.3. Student responses. The reported aggregate scores across six items were 321 before media use and 680 afterward for 27 students. Dividing these totals by 27×6 yields mean item responses of 1.98 and 4.20, respectively. With a maximum item score of 5, the normalized change is $(4.20 - 1.98)/(5 - 1.98) \approx 0.73$. Although normalized gain was originally developed to describe changes in performance on conceptual tests [33], it is used here only as a descriptive index of change in questionnaire scores. In the absence of the six item texts and evidence of their measurement properties, the result cannot be interpreted as a verified gain in mathematical knowledge.

$$g = \frac{Sf - Si}{5 - Si} \quad (3)$$

2.7. Ethical Considerations

The original account reports school authorization, voluntary participation, informed consent, and anonymization of personal data. Because students were minors and the example media could transmit named quiz scores to Canva Sheet, the description is insufficient to verify parent or guardian permission, student assent, access controls, data retention periods, or platform-specific safeguards. No additional consent or data-protection procedure is claimed.

3. Results and Discussion

3.1. Result

3.1.1. Development of the AI Prompt Blueprint

The study reports the development of 14 prompt blueprints covering lines and angles, plane geometry, and statistics. Each was organized around Role, Context, Task, Constrain, and Output (Figure 2) and revised through an iterative generation-and-review process (Figure 3). One statistics example is documented in detail, but the titles, revision histories, and individual outcomes of the remaining 13 prompts are not reported sufficiently to reconstruct the complete development process. The documented example and total number of products are therefore presented without implying that all 14 items underwent identical review.

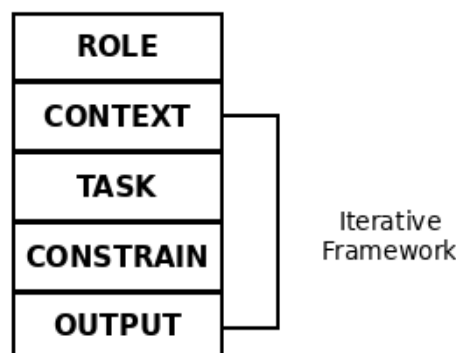


Figure 2. RCTCO prompt structure used in the development process.

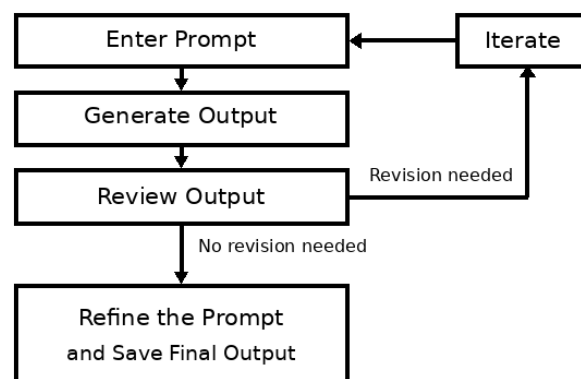


Figure 3. Reported prompt-generation, reflection, and revision cycle.

In the statistics example, the AI was assigned the roles of web developer and mathematics teacher. The prompt specified learning content on samples, populations, frequency distributions, measures of central tendency, and statistical graphs, as well as a visual calculator, ten quiz items, tooltips, feedback, and responsive design. Table 1 summarizes these requirements. The example has been expressed in academic English without introducing additional development activities.

Table 1. Example of the RCTCO blueprint for Grade 7 statistics

Blueprint component	Specification
Role	Web developer and Grade 7 mathematics teacher.
Context	Seventh-grade statistics, including populations, samples, frequency distributions, and measures of centre and spread.
Task	Develop a dashboard, learning-content pages, a visual statistics calculator, and a ten-question quiz. Use Canva Sheet for recording quiz information; provide a clear, high-contrast interface, tooltips, and response feedback.
Constrain	Check mathematical terms, calculations, and representations; do not assume that generated content is automatically correct.
Output	Responsive web-based mathematics learning media.
Example prompt	Act as a web developer and Grade 7 mathematics teacher. Create interactive media for the statistics topic. Include learning objectives covering samples and populations, frequency tables, bar/histogram and pie-chart representations, mean, median, mode, and range. Provide a dashboard headed "Statistics and Data Representation", three navigation destinations (materials, visual calculator, quiz), ten practice questions, tooltips, and immediate feedback. Include entry of student name and class before quiz scores are sent to Canva Sheet, as specified in the original prompt. Use a light, readable interface that adapts to desktop and smartphone displays. Check every formula, answer, and diagram before publication.

Table 2. Documented initial and first revision of the statistics prompt

Iteration	Instruction	Reported outcome
0	Initial broad prompt covering the required topic and main interface elements.	A usable initial medium was generated; the original text described few errors but gave no error log.
1	Add interactive population estimation; strengthen frequency-table and graph illustrations; provide ten randomized examples for mean, median and mode with worked solutions.	The revised media incorporated further visualization functions and explanatory elements; no quantitative improvement metric was reported.

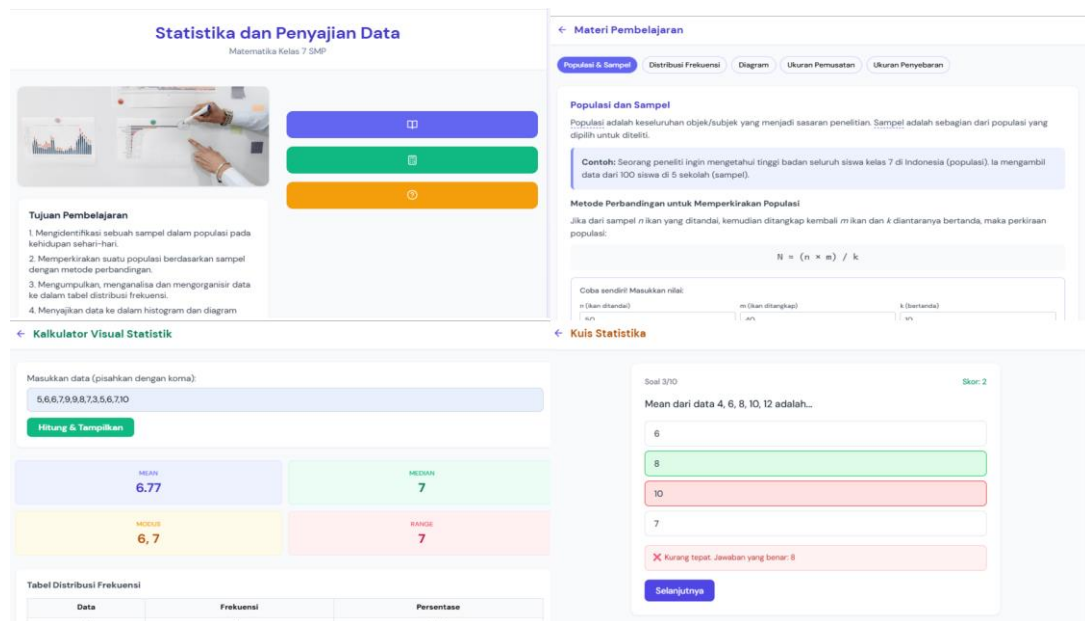


Figure 4. Example of statistics learning media generated with Canva Code.

3.1.2. Expert Ratings and Participant Responses

An expert panel evaluated both the prompt blueprints and the generated media, although the instruments contained different numbers of items for different specialties. Tables 3 and 4 retain the reported assessor totals. Normalized summary scores of approximately 0.90 for prompts and 0.91 for media indicate favorable judgments within this evaluation. The asterisks in Tables 3 and 4 identify values originally reported as Aiken's V. They can be reproduced from the available totals as normalized mean assessor ratings but cannot be verified as item-level validity coefficients.

Teachers gave favorable descriptive ratings on the questionnaire (Table 5), while students' aggregate scores on six questionnaire items increased between administrations (Table 6). These results are descriptive. One reported student had identical scores before and after media use (20 in both assessments), indicating that responses did not increase for every participant. The previously reported teacher kappa of 0.91 is retained in Table 5 with an asterisk for transparency, but the categorical agreement data required for verification are unavailable. Likewise, the asterisk in Table 6 identifies a change in questionnaire scores, not a demonstrated gain in mathematics achievement. Complete item-level records would be needed to describe variation across all 27 students.

Table 3. Prompt blueprint expert ratings (original scores retained)

Assessor role	Items	Total score	Mean item	per	Mean - 1
Media specialist	20	93	4.650		3.650
Media specialist	20	91	4.550		3.550
Media specialist	20	97	4.850		3.850
Mathematics specialist	16	72	4.500		3.500
Language specialist	8	36	4.500		3.500
Reported: 0.90*			$V = \frac{18.02}{5(5 - 1)}$ V = 0.90		

Table 4. Generated media expert ratings (original scores retained)

Assessor role	Items	Total score	Mean per item	Mean - 1
Media specialist	21	98	4.667	3.667
Media specialist	21	90	4.286	3.286
Media specialist	21	102	4.857	3.857
Mathematics specialist	23	113	4.913	3.913
Language specialist	8	36	4.500	3.500
Reported: 0.91*		$V = \frac{18.16}{5(5 - 1)}$ V = 0.91		

Table 5. Descriptive teacher questionnaire ratings

Teacher	Number of items	Observed total	Mean (max = 5)
Teacher 1	22	101	4.591
Teacher 2	22	99	4.500
Teacher 3	22	104	4.727
Reported 0.91*	κ:	$\kappa = \frac{0.92 - 0.078}{1 - 0.078}$ K = 0.91	

Table 6. Student six-item questionnaire scores before and after media use

Student code	Before (6 items)	After (6 items)
S01	13	25
S02	20	20
...	...	...
S27	12	27
Total (n=27)	321	680
Mean per item	1.9815	4.1975
Score change: $g \approx 0.73^*$		$g = \frac{4.20 - 1.98}{5 - 1.98}$ g = 0.73

The available evidence supports three qualified findings: (1) the study produced a structured RCTCO prompt framework and reported 14 prompt designs; (2) five expert assessors gave favorable aggregate ratings to the prompts and generated media; and (3) participating teachers provided high questionnaire ratings, while students' mean six-item questionnaire scores increased in this single-school evaluation. These findings do not, by themselves, demonstrate error-free mathematical content, independent media production by teachers, or a causal improvement in academic achievement.

3.2. Discussion

3.2.1. Interpretation of Prompt and Media Ratings

Evaluating an AI instruction separately from the instructional resource it generates is an important methodological distinction in instructional design. The favorable normalized expert ratings (0.90 for prompts and 0.91 for media) suggest that the RCTCO workflow was feasible in this local development context. Nevertheless, the proximity of the two scores does not establish that prompt quality caused media quality, particularly because the aggregated ratings came from different item sets. Research on educational prompt engineering similarly emphasizes iterative specification, contextualized tasks, and human evaluation [17], [19], [21], [24]. Future validation should report item-

level expert judgments and assess generated mathematical content against documented reference solutions.

3.2.2. Teacher Response and Practical Implications

The three mathematics teachers rated the media favorably, with a mean questionnaire score of 4.61/5. This suggests a positive response among the participating teachers, but it does not establish that educators without programming experience could produce media independently or that the reported kappa represented genuine inter-rater agreement. Research on educators' adoption of AI highlights the importance of AI literacy, professional support, institutional conditions, and task-technology fit [34]–[39]. The practical implication is to regard the blueprint as a potentially useful scaffold for teacher-led design and review, rather than as a self-validating substitute for professional instructional judgment [5], [23].

3.2.3. Student Questionnaire Change and Learning Theory

Mean responses to the six student statements rose from 1.98 to 4.20, corresponding to a normalized questionnaire-score change of approximately 0.73. This increase indicates more favorable responses following media use, although the construct measured remains uncertain because the item wording was not provided. The findings therefore cannot establish a causal improvement in mathematics proficiency. Multimedia learning theory nevertheless offers a rationale for examining visual representations and immediate feedback [8], [9]. Controlled studies of GenAI-supported mathematics learning indicate that instructional design and safeguards can shape outcomes while highlighting potential accuracy problems [13]–[16], [27]. These studies provide relevant context but do not replace evidence from the present sample.

3.2.4. Scope of Contribution and Comparison with Prior Work

The study offers three bounded contributions to the development of instructional media:

- 3.2.4.1. First, it specifies the Role–Context–Task–Constrain–Output structure for generating interactive mathematics media. This complements research on prompt engineering and instructional design without establishing RCTCO as a universally effective model [18], [22], [26], [28], [29].
- 3.2.4.2. Second, it treats prompts and generated media as separate objects of expert review, rather than assuming that a functional web interface is sufficient evidence of instructional quality. This distinction is consistent with calls for critical evaluation of AI-produced teaching materials [6], [25], [40].
- 3.2.4.3. Third, it documents iterative development of a Grade 7 statistics resource using Canva Code. This application differs in focus from previous evaluations of Canva-based learning materials [10]–[12] and AI-adaptive learning systems [41]. However, the study does not establish superiority over these approaches because no direct comparison was performed.

3.2.5. Limitations

The following limitations constrain the interpretation and transferability of the findings:

- 3.2.5.1. First, the study involved a single school, three mathematics teachers, and 27 students. The random sampling procedure was not documented in sufficient detail, limiting generalizability.
- 3.2.5.2. Second, the products covered only three clusters of Grade 7 mathematics content. Only one prompt revision sequence is documented in detail, rather than the development histories of all 14 products.

- 3.2.5.3. Third, the assessment instruments and item-level records were unavailable. Consequently, the reported summary scores cannot be independently verified as Aiken's V or Cohen's kappa, and the internal consistency of the student questionnaire remains unknown.
- 3.2.5.4. Fourth, the immediate, single-group pre/post questionnaire design included neither a comparison group nor a delayed retention assessment. The six statements were not established as objective measures of mathematics achievement.
- 3.2.5.5. Fifth, the study did not report systematic analyses of mathematical error rates, AI hallucinations, accessibility, or performance across Canva Code updates. Human verification therefore remains necessary [3], [4], [7], [42].

3.2.6. Directions for Further Research

Future research should document all 14 prompt revision histories, report item-level expert evaluations, and establish the validity of measurement instruments before conducting larger studies. Multi-school investigations with clearly documented sampling and suitable comparison groups would better assess mathematics achievement. Independent teacher-authoring tasks could directly evaluate whether educators without programming backgrounds can produce media autonomously. Comparative studies of prompting approaches, data transparency and protection, retention, and platform stability would also address important limitations. Research combining iterative design with carefully verified learning processes provides useful methodological guidance [5], [20], [24], [29].

4. Conclusion

Using the R&D approach and ADDIE model, this study developed a structured RCTCO blueprint and reported 14 Canva Code prompts for Grade 7 mathematics learning media. Five experts assigned normalized summary ratings of approximately 0.90 to the prompts and 0.91 to the generated media. Three teachers gave a mean questionnaire response of 4.61/5, while students' mean responses to six questionnaire statements increased from 1.98 to 4.20 (normalized change ≈ 0.73). Because item-level validation records, categorical inter-rater data, and evidence that the six statements measured mathematics achievement were unavailable, these results cannot be treated as verified Aiken's V, Cohen's kappa, or causal gains in mathematical knowledge. The principal contribution is a documented development process that evaluates prompt instructions separately from the generated instructional media. Stronger claims will require complete instrument documentation, independent mathematical-content review, more rigorous evaluation designs, and appropriate safeguards for student data.

Declarations

Author Contributions

Habib Burrahman contributed to the conceptualization, research design, data collection, data analysis, and preparation of the initial manuscript. Firdaus Annas contributed to supervision, methodological validation, critical revision, and approval of the final manuscript, as stated in the original authorship declaration.

Funding

The authors reported no specific external funding; study expenses were self-funded.

Data Availability

The original data-availability statement described the dataset and prompt blueprints as both publicly available and accessible upon request. Because no public repository was identified, unrestricted availability cannot be confirmed. Requests for de-identified supporting records and prompt artifacts may be directed to the corresponding author, subject to the school's and participants' privacy requirements. The authors must confirm the precise access conditions before publication.

Conflicts of Interest

The authors declared no financial or non-financial conflicts of interest relevant to this work.

Generative AI Disclosure

Canva Code was the generative AI platform used to develop the instructional materials investigated in this study. The original manuscript contained three mutually inconsistent options regarding the separate use of Qwen 3.7-Plus for manuscript preparation. The authors must confirm the applicable option and retain only an accurate disclosure before submission.

Additional Information

Correspondence and requests for research materials may be addressed to Habib Burrahman using the email addresses above.

ORCID

[Habib Burrahman],  <https://orcid.org/0009-0007-1536-6557>

References

- [1] Z. Bahroun, C. Anane, V. Ahmed, and A. Zacca, "Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis," *Sustainability*, vol. 15, no. 17, Art. no. 12983, 2023, doi: 10.3390/su151712983.
- [2] T. K. F. Chiu, Q. Xia, X. Zhou, C. S. Chai, and M. Cheng, "Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 4, Art. no. 100118, 2023, doi: 10.1016/j.caeai.2022.100118.
- [3] X. Wang, Z. Zainuddin, and C. H. Leng, "Generative artificial intelligence in pedagogical practices: A systematic review of empirical studies (2022–2024)," *Cogent Education*, vol. 12, no. 1, Art. no. 2485499, 2025, doi: 10.1080/2331186X.2025.2485499.
- [4] A. Yusuf, N. Pervin, M. Román-González, and N. M. Noor, "Generative AI in education and research: A systematic mapping review," *Review of Education*, vol. 12, no. 2, Art. no. e3489, 2024, doi: 10.1002/rev3.3489.
- [5] Y. H. Cheah, J. Lu, and J. Kim, "Integrating generative artificial intelligence in K-12 education: Examining teachers' preparedness, practices, and barriers," *Computers and Education: Artificial Intelligence*, vol. 8, Art. no. 100363, 2025, doi: 10.1016/j.caeai.2025.100363.
- [6] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, "A SWOT analysis of ChatGPT: Implications for educational practice and research," *Innovations in Education and Teaching International*, vol. 61, no. 3, pp. 460–474, 2024, doi: 10.1080/14703297.2023.2195846.
- [7] F. Miao and W. Holmes, *Guidance for generative AI in education and research*. UNESCO, 2023. [Online]. Available: <https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research>
- [8] J. G. Cromley and R. Chen, "A meta-analysis of Richard Mayer's multimedia learning research: Searching for boundary conditions of design principles across multiple media types," *Educational Research Review*, Art. no. 100730, 2025, doi: 10.1016/j.edurev.2025.100730.
- [9] R. E. Mayer, "The past, present, and future of the cognitive theory of multimedia learning," *Educational Psychology Review*, vol. 36, Art. no. 8, 2024, doi: 10.1007/s10648-023-09842-1.
- [10] J. M. Asih and R. Yulastuti, "Development of Canva-based interactive mathematics learning media to improve learning outcomes of SMP students," *EMTEKA: Jurnal Pendidikan Matematika*, vol. 6, no. 2, pp. 735–746, 2025, doi: 10.24127/emteka.v6i2.9553.
- [11] B. Malo, M. H. S. Raja, K. Nona, F. Sizi, and R. Nembo, "Pemanfaatan Canva sebagai media pembelajaran yang kreatif dan interaktif dalam meningkatkan hasil belajar matematika SMP," *JagoMIPA: Jurnal Pendidikan Matematika dan IPA*, vol. 5, no. 3, pp. 942–952, 2025, doi: 10.53299/jagomipa.v5i3.2088.
- [12] N. M. Sumartiwi and P. R. Ujianti, "Media pembelajaran audio visual berbasis aplikasi Canva pada materi keliling dan luas lingkaran," *Jurnal Pedagogi dan Pembelajaran*, vol. 5, no. 2, pp. 220–230, 2022, doi: 10.23887/jp2.v5i2.47626.
- [13] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakçı, and R. Mariman, "Generative AI without guardrails can harm learning: Evidence from high school mathematics," *Proceedings of the National Academy of Sciences*, vol. 122, no. 26, Art. no. e2422633122, 2025, doi: 10.1073/pnas.2422633122.
- [14] B. Liu, W. Zhang, and F. Wang, "Can generative artificial intelligence effectively enhance students' mathematics learning outcomes? A meta-analysis of empirical studies from 2023 to 2025," *Education Sciences*, vol. 16, no. 1, Art. no. 140, 2026, doi: 10.3390/educsci16010140.
- [15] Z. A. Pardos and S. Bhandari, "ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills," *PLOS ONE*, vol. 19, no. 5, Art. no. e0304013, 2024, doi: 10.1371/journal.pone.0304013.
- [16] W. Xing, Y. Song, C. Li, Z. Liu, W. Zhu, and H. Oh, "Development of a generative AI-powered teachable agent for middle school mathematics learning: A design-based research study," *British Journal of Educational Technology*, vol. 56, no. 5, pp. 2043–2077, 2025, doi: 10.1111/bjet.13586.
- [17] D. Lee and E. Palmer, "Prompt engineering in higher education: A systematic review to help inform curricula," *International Journal of Educational Technology in Higher Education*, vol. 22, Art. no. 7, 2025, doi: 10.1186/s41239-025-00503-7.

- [18] Y. Qian, "Pedagogical applications of generative AI in higher education: A systematic review of the field," *TechTrends*, vol. 69, pp. 1105–1120, 2025, doi: 10.1007/s11528-025-01100-1.
- [19] Y. Qian, "Prompt engineering in education: A systematic review of approaches and educational applications," *Journal of Educational Computing Research*, vol. 63, no. 7–8, pp. 1782–1818, 2025, doi: 10.1177/07356331251365189.
- [20] H. Thomann and V. Deutscher, "Scaffolding through prompts in digital learning: A systematic review and meta-analysis of effectiveness on learning achievement," *Educational Research Review*, vol. 47, Art. no. 100686, 2025, doi: 10.1016/j.edurev.2025.100686.
- [21] J. L. Araújo and I. Saúde, "Can ChatGPT enhance chemistry laboratory teaching? Using prompt engineering to enable AI in generating laboratory activities," *Journal of Chemical Education*, vol. 101, pp. 1858–1864, 2024, doi: 10.1021/acs.jchemed.3c00745.
- [22] R. O. Davis and Y. J. Lee, "Prompt: ChatGPT, create my course, please!", *Education Sciences*, vol. 14, no. 1, Art. no. 24, 2024, doi: 10.3390/educsci14010024.
- [23] A. ElSayary, M. A. Kuhail, and Z. Hojeij, "Examining the role of prompt engineering in utilizing generative AI tools for lesson planning: Insights from teachers' experiences and perceptions," *Human Behavior and Emerging Technologies*, vol. 2025, Art. no. 9986139, 2025, doi: 10.1155/hbe2/9986139.
- [24] B. Hu, L. Zheng, J. Zhu, L. Ding, Y. Wang, and X. Gu, "Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1445–1459, 2024, doi: 10.1109/TLT.2024.3384765.
- [25] M. Moundridou, N. Matzakos, and S. Doukakis, "Generative AI tools as educators' assistants: Designing and implementing inquiry-based lesson plans," *Computers and Education: Artificial Intelligence*, vol. 7, Art. no. 100277, 2024, doi: 10.1016/j.caeai.2024.100277.
- [26] M. Brasken, K.-E. Berts, S. Wallin, and S. Frilund, "Students' use of ChatGPT in an algebra class: A case study of prompts and attitudes," *LUMAT: International Journal on Math, Science and Technology Education*, vol. 13, no. 1, 2025, doi: 10.31129/LUMAT.13.1.2877.
- [27] W. McGalliard and S. Otten, "AI responses to challenging problems and educator responses to AI availability," *Digital Experiences in Mathematics Education*, vol. 11, pp. 319–332, 2025, doi: 10.1007/s40751-024-00167-4.
- [28] I. Celik, S. Kontkanen, J. Laru, and A. A. Dalyanci, "Co-constructing adaptive lesson plans with GenAI: Pre-service teachers' Intelligent-TPACK and prompt engineering strategies," *Computers & Education*, vol. 241, Art. no. 105485, 2026, doi: 10.1016/j.compedu.2025.105485.
- [29] Y. Sun and Y. Huang, "Exploring the role of generative AI in collaborative lesson planning for pre-service teachers," *Journal of Computer Assisted Learning*, vol. 41, no. 4, Art. no. e70098, 2025, doi: 10.1111/jcal.70098.
- [30] R. M. Branch, *Instructional design: The ADDIE approach*. Springer, 2009, doi: 10.1007/978-0-387-09506-6.
- [31] L. R. Aiken, "Three coefficients for analyzing the reliability and validity of ratings," *Educational and Psychological Measurement*, vol. 45, no. 1, pp. 131–142, 1985, doi: 10.1177/0013164485451012.
- [32] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.
- [33] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American Journal of Physics*, vol. 66, no. 1, pp. 64–74, 1998, doi: 10.1119/1.18809.
- [34] A. Hazzan-Bishara, O. Kol, and S. Levy, "The factors affecting teachers' adoption of AI technologies: A unified model of external and internal determinants," *Education and Information Technologies*, vol. 30, pp. 15043–15069, 2025, doi: 10.1007/s10639-025-13393-z.
- [35] M. Kokoç, Ş. Ağırman, and B. N. Yağmurlugil, "Mapping in-service teacher AI literacy: A systematic review of empirical studies," *Teaching and Teacher Education*, vol. 181, Art. no. 105707, 2026, doi: 10.1016/j.tate.2026.105707.
- [36] S. C. Kong, Y. Yang, and C. Hou, "Examining teachers' behavioural intention of using generative artificial intelligence tools for teaching and learning based on the extended technology acceptance model," *Computers and Education: Artificial Intelligence*, vol. 7, Art. no. 100328, 2024, doi: 10.1016/j.caeai.2024.100328.
- [37] Z. Song, J. Qin, F. Jin, W. M. Cheung, and C.-H. Lin, "A case study of teachers' generative artificial intelligence integration processes and factors influencing them," *Teaching and Teacher Education*, vol. 165, Art. no. 105157, 2025, doi: 10.1016/j.tate.2025.105157.
- [38] J. Zhao and X. Yang, "Understanding K-12 teachers' adoption of generative AI through an integrated technology acceptance and perceived risk framework," *Discover Education*, vol. 5, Art. no. 592, 2026, doi: 10.1007/s44217-026-01611-w.
- [39] Y. Zhu, H. Jiang, and R. Chugh, "Empowering STEM teachers' professional learning through GenAI: The roles of task-technology fit, cognitive appraisal, and emotions," *Teaching and Teacher Education*, vol. 167, Art. no. 105204, 2025, doi: 10.1016/j.tate.2025.105204.
- [40] D. Hauk and N. Soujon, "How reliable are large language models in analyzing the quality of written lesson plans? A mixed-methods study from a teacher internship program," *Computers and Education*:

- Artificial Intelligence*, vol. 10, Art. no. 100538, 2026, doi: 10.1016/j.caeai.2025.100538.
- [41] Rustiyana, P. Wapun, J. Lima, and H. A. Makie, "AI-based adaptive learning systems to support the Merdeka Curriculum: Personalizing mathematics learning at the junior high school level," *Journal International Inspire Education Technology*, vol. 4, no. 3, pp. 262–281, 2025, doi: 10.55849/jiiet.v4i3.1085.
- [42] G. Cooper, "Examining science education in ChatGPT: An exploratory study of generative artificial intelligence," *Journal of Science Education and Technology*, vol. 32, pp. 444–452, 2023, doi: 10.1007/s10956-023-10039-y.