

Received : 13-03-2026

Accepted : 28-04-2026

Examining the Impact of Structured Prompt Frameworks on the Pedagogical Quality and Cognitive Level of AI-Generated Research Methodology Lesson Plans

Athina BOUKHELOUFBatna 2 University, Algeria, a.boukhelouf@univ-batna2.dz

Abstract

This study uses a 2x2 factorial design to examine the effects of AI model (ChatGPT vs Copilot), prompt framework (CRISPE vs TAG) and their interaction on the pedagogical quality and cognitive demand of lesson plans. Twenty research methodology lesson plans were generated under the four conditions and evaluated using an analytic evaluation rubric. The latter was designed and validated to assess five major dimensions, namely learning objectives and pedagogical alignment, lesson structure and clarity, cognitive demand and higher-order thinking skills, practical usability, learner engagement and motivation. The two-way analysis of variance (ANOVA) indicates that both AI model and prompt framework significantly influence the pedagogical quality of research methodology lesson plans. Lesson plans generated using CRISPE achieved substantially higher pedagogical quality scores than those generated by TAG, across both AI models. Particularly, lesson plans produced by ChatGPT under the CRISPE framework proved to be well-structured and pedagogically aligned. On the other hand, TAG lesson plans revealed multiple weaknesses relevant to alignment, clarity and learner engagement. Across all conditions, lesson plans failed to account for higher order thinking skills consistently. Overall, this apparent gap implies that AI lesson plans cannot serve as pedagogical alternatives to human lesson plans. Indeed, human pedagogical expertise is crucial to ensure that AI lesson plans are fully consistent with pedagogy, cognitive rigor and education standards.

Keywords: prompt framework, pedagogical quality, cognitive demand, higher order thinking skills, Chat GPT, Copilot

1. Introduction

The advancement of Artificial Intelligence (AI) has drastically transformed various life domains, and education is no exception. Indeed, the implementation of AI tools in education has the potential to revolutionize and optimize the whole process of teaching and learning. Particularly, AI has introduced innovative pedagogical tools that use both virtual and augmented reality (Chen et al., 2020). In other words, AI tools can help both teachers and learners enhance the way they teach and learn by creating immersive, adaptive, and engaging environments.

The integration of AI Large Language Models (LLMs), such as ChatGPT and Copilot, into education offers several benefits. It assists teachers in designing and tailoring personalized courses that suit different learning styles and levels (Fang & Broussard, 2024; Holmes et al., 2019). Especially, AI affords diversified opportunities for educators to develop high-quality instructional materials and effective lesson plans that align with curriculum objectives (Fang & Broussard, 2024). Such advantages can be the outcome of entering a well-structured prompt into AI-based tools. That is to say, by inputting a specific prompt, LLMs can design and format a variety of well-structured lesson plans and innovative materials efficiently. However, having the capacity to generate well-structured lesson plans and instructional materials does not necessarily equate to pedagogical soundness and rigor (Kasneci et al., 2023). In this regard, recent studies accentuate that the pedagogical soundness of AI-generated lesson plans remains questionable (Karaman & Goksu, 2024; Powell & Courchesne, 2024). They emphasize multifaceted concerns relevant to accuracy, contextual adaptability, cognitive engagement, and pedagogical applicability. Consequently, there is a need for empirical research to critically evaluate the quality and effectiveness of AI-generated lesson plans from both pedagogical and cognitive perspectives. This would provide evidence on the key factors influencing AI output, thereby guiding its effective integration into the teaching process. In this vein, this study aims to examine the overall pedagogical quality of research methodology lesson plans generated by ChatGPT and Copilot. Specifically, it assesses the extent to which prompt structure and AI model can influence the evaluation metrics, including learning objectives and pedagogical alignment, lesson structure and clarity, cognitive demand and higher order thinking skills, practical usability, learner engagement and motivation across the generated lesson plans. Accordingly, the following questions and hypotheses are formulated to guide this investigation:

1. What is the overall pedagogical quality of AI-generated research methodology lesson plans?
2. To what extent do prompt frameworks (CRISPE vs. TAG) influence the pedagogical quality of AI-generated research methodology lesson plans?
3. To what extent do AI model (Chat GPT vs Copilot) affect the pedagogical quality of AI-generated research methodology lesson plans?
4. To what degree are higher-order thinking skills present in AI-generated research methodology lesson plans produced under different prompt frameworks?

5. Does the interaction between AI models and prompt frameworks determine the pedagogical quality of research methodology lesson plans?

Hypotheses:

H1: There is a significant difference in the pedagogical quality score of research methodology lesson plans generated by Chat GPT and Copilot.

H2: There is a significant difference in the pedagogical quality scores between research methodology lesson plans produced under the CRISPE framework and those produced under the TAG framework.

H3: There is a significant interaction effect between AI model and prompt framework on the pedagogical quality of research methodology lesson plans.

To ground this study theoretically and empirically, the following section reviews the relevant literature.

2. Literature Review:

Generative AI (Gen AI) introduced new paradigms to education and teaching, in general, and lesson planning, in particular. In fact, it can offer limitless potential and adaptable lesson plans at an accelerated pace. This may appeal to many educators and teachers as it reduces planning time and cognitive load while generating various innovative ideas.

Several studies maintain that AI-generated lesson plans can be as effective as those designed by humans (Karaman & Goksu, 2024; Fan et al., 2024). They highlight that AI-generated lesson plans can be pedagogically sound. Also, they can improve students' learning outcomes and achievement. Nonetheless, these advantages are strongly contingent on the structure of the input and output editing. Indeed, AI models are incapable of generating content that conforms to the pedagogical norms unless properly guided (Dornburg & Davin, 2025). In other words, prompt structure plays a crucial role in improving the pedagogical quality and consistency of AI-generated lesson plans.

In this vein, researchers developed various frameworks to guide AI models in producing meaningful output. Six categories of prompt patterns (White et al., 2023), CLEAR framework (Lo, 2023), AIPROMT (Korynski et al., 2023) and CRISPE (Open AI) are notable examples. These frameworks provide structured and strategic guidelines to derive potential benefits when engaging with AI.

For instance, CRISPE denotes the Capacity and Role, Insight, Statement, Personality and Experiment (Wang et al.2021). Thus, to generate effective AI outputs, users should specify the role of the AI model. Thereafter, they should provide information relevant to the context. Next, users should articulate what they want from the AI model exactly, specifying the style and manner they prefer. Finally, they should request the AI model to produce multiple answers. Respecting the aforementioned guidelines is likely to generate comprehensive and significant outputs (Wang et al.2021). However, users should be able to critically assess the reliability of the output. Therefore, background knowledge and critical

mind are pivotal when engaging with AI. In the educational context, teachers and educators should not only develop prompt engineering skills for optimum benefits. They are required to be discerning and insightful. Indeed, they should critically evaluate AI-generated lesson plans before implementation, as some generated lesson plans may appear well-structured, albeit limited.

In other words, AI efficiency in generating well-structured and coherent lesson plans may not correlate with effectiveness. In this regard, some researchers like Baytak (2024), Kahraman and Kiyici (2025) evaluated AI-generated lesson plans and delineated various limitations. These flaws pertain to curricular alignment, applicability, adaptability, fabricated resources and factual errors (Powell & Courchesne, 2024). According to Baytak (2024), AI-generated lesson plans fall short of meeting the international curricular standards and objectives. Besides, he reported that these plans failed to meaningfully integrate technology strategies.

Conversely, Lee and Zhai (2024) pinpoint that AI lesson plans demonstrated a moderate alignment with curricula standards. Particularly, they emphasize the fact that these lesson plans are generic and context-insensitive. Hence, AI-generated lesson plans require adaptation to suit the needs of the local classrooms.

Likewise, Kahraman and Kiyici (2025) posit that AI-generated lesson plans can adequately address the learning objectives. However, they are devoid of formative assessment practices and reflective components. Also, they insufficiently consider various learning needs and abilities.

Moreover, AI-generated lesson plans may contain inaccurate information (Powell & Courchesne, 2024). Factual errors and fabricated sources or results can be serious risks and negative repercussions of using AI lesson plans without assessing the reliability of their content, especially when the subject demands accuracy like mathematics, physics, science, and research methodology. Therefore, it is mandatory for teachers to examine and identify factual inconsistencies and hallucinated content, if any. This will ensure proper and effective practices. In addition, it will promote ethical responsibility.

Furthermore, Liu's study (2025) points out an additional limitation concerning AI-generated lesson plans and intellectual demand. The study evaluated the cognitive rigor of fifteen AI-generated lessons' learning objectives. The findings indicate that all learning objectives revolve around the lower tiers of Bloom's taxonomy. In other words, recalling and understanding are by default the predominant targeted cognitive skills across all lesson plans. Inevitably, this will limit instructional depth and reflection among students. This gap underscores the necessity of human expertise to ensure that lessons plans consistently stimulate higher-order thinking, namely analysis, evaluation and creation.

In conclusion, research findings revealed significant prospects and gaps in empirical and multidimensional evaluation of AI-generated lesson plans. Although AI models are capable of efficiently producing well-structured and coherent lesson plans, human guidance and expertise are imperative to guarantee that these AI-generated lesson plans are fully aligned with the set standards and applicable to a particular educational context.

3. Methodology Design

This study used a balanced 2x2 factorial design to assess the effects of two independent variables namely AI model (Chat GPT vs Copilot), prompt framework (CRISPE vs TAG), and their interaction, on the dependent variable: the pedagogical quality of research methodology lesson plans. It opted for a validated quantitative analytic rubric to evaluate five metrics: learning objectives and pedagogical alignment, cognitive demand, higher-order thinking skills, lesson structure and clarity, practical usability and learner engagement and motivation.

3.1 Lesson Topic

The research methodology topics selected for the prompts were Quantitative Research, Qualitative Research, Research Problem, Research Hypothesis, and Data Gathering Tools. These topics form an integral part of the curriculum meant to introduce 3rd year Algerian EFL students to research methodology.

3.2 Prompt structure

To examine how the structure of the input may influence the quality of AI-generated lesson plans, two prompt frameworks were used. Information pertaining to duration (60 minutes), level (3rd year LMD students of English as a Foreign Language), objectives, and topic remained identical across all prompts to ensure impartiality. Table 01 illustrates prompts design using TAG and CRISPE frameworks. These prompts were submitted to Chat GPT and Copilot to generate lesson plans on Research Problem. The same process was undergone to produce lesson plans on the aforementioned topics.

Table 01 Prompts Design Example

CRISPE: Capacity and Role-Insight-Statement- Personality-Experiment.	TAG: Task- Audience – Goal
Capacity: You are an assistant with expertise in both Research Methodology and EFL Instruction at the university level. Role: Act as a university lecturer and prepare a 60 minutes lesson plan for third year Algerian EFL students on Research Problem. Insight: • The students have basic knowledge of research concepts. • They learn best through interactive, scaffold and task-based activities. • The lesson should help	Task: Generate a complete/engaging 60 minutes lesson plan on Research Problem. Audience: 3rd Algerian EFL students. Goal: enable students to formulate adequate research problems.

students formulate adequate research problems. Statement: design a full 60-minutes lesson plan that includes objectives, warm up activities, assessment, materials, time allocation and homework. Personality: The lesson should be engaging, practical and easy to implement. Experiment: Provide 3 alternative versions of the lesson plan: Each version should meet the same learning objectives.	
---	--

3.3 Data Collection

CRISPE and TAG were used to design prompts to guide Chat GPT and Copilot in generating lesson plans for the topics described in section 3.1. For each topic, four lesson plans were produced (one plan per topic, AI model and prompt framework). Therefore, the result was 20 lesson plans organized in 4 experimental cells. This permits obtaining a balanced factorial design of four equal cells ($n = 5$ per condition), thereby strengthening the robustness of the analysis. Indeed, balanced factorial designs are argued to strengthen the robustness of the two-way ANOVA, even with modest sample sizes (Field, 2013; Kirk, 2013).

It is noteworthy that all these plans were anonymized and stored in text format. Then they were printed along with evaluation rubrics, and administered to a panel of 3 experts in research methodology (raters).

3.4 Evaluation Rubric

Grounded in Biggs' constructive alignment model and Bloom's taxonomy, the evaluation rubric was designed, piloted and validated by nine teachers, achieving a content validity index above 0.90. This rubric aims to systematically assess the following metrics using a 5-point Likert scale:

- Learning Objectives and Pedagogical Alignment
- Lesson Structure and Clarity
- Cognitive demand and Higher Order Thinking skills
- Practical Usability
- Learner Engagement and Motivation

3.4.1. *Learning objectives and pedagogical alignment*: the first metric evaluates the major determinants of effective lesson plans like objective clarity, curriculum alignment, assessment integration and instructional strategy.

3.4.2. *Lesson structure and clarity*: the second metric examines the extent to which elements of the lesson are logically sequenced and coherent. Additionally, it assesses the clarity of instructions and time allotted for each activity.

3.4.3. *Cognitive Demand and Higher-Order thinking Skills*: the third metrics uses Bloom's taxonomy as framework to assess the degree to which the proposed objectives, activities and instructions encourage higher-order thinking skills and promote reasoning, reflection and problem-solving. Ultimately, it scrutinizes the cognitive complexity of tasks.

3.4.4. *Practical Usability*: the fourth metric evaluates the accuracy, accessibility and applicability of the suggested materials and content. Also, it examines whether the lesson plans can be adapted to different contexts or students' level. Further, it deals with the clarity and sufficiency of the guidance.

3.4.5. *Learner engagement and motivation*: the fifth metric determines the potential of the lesson plan to motivate and engage students through analyzing the type of tasks.

The previously explained metrics allowed the study to ascertain whether AI-generated lesson plans are pedagogically effective. They ensure objective, quantitative and replicable evaluation that permits comparison of AI-generated lesson plans across different models and prompt framework (See Table 2).

Table 02 Metrics and scoring range

Metrics	Score (Min-Max)
Learning Objectives and Pedagogical Alignment	4-20
Lesson Structure and Clarity	3-15
Cognitive demand and Higher Order Thinking skills	3-15
Practical Usability	3-15
Learner Engagement and Motivation	2-10
Total	15-75

3.5. Inter-rater Reliability Procedures

To minimize subjective bias and ensure the robust assessment of AI- generated lesson plans, Intraclass Correlation Coefficient (ICC) was employed. It measures the consistency of agreement among raters (Shrout & Fleiss, 1979). Especially, two-way mixed model and absolute agreement for three raters were selected. The results indicate excellent reliability, ICC (3, K) = .995, CI [.989,.998], $p < .001$. They imply that there is a high degree of agreement among raters. Additionally, they validate ratings and justify the use of averaged scores.

3.6. Data Analysis

The analytical framework of this study involves two complementary approaches. Descriptive statistics were used to comprehensively summarize the dataset and identify its characteristics. Especially, the averaged mean score of the three raters for each experimental condition was calculated. Subsequently, the normality and homogeneity of the results were tested through Levene's test. After meeting the fundamental statistical assumptions, two-way analysis of variance ANOVA was conducted to examine the effect of AI model and prompt framework, as well as their interaction, on research methodology lesson plans pedagogical quality. This factorial design permits the simultaneous examination of the independent contribution of each factor as well as their interaction effect.

5.Results

5.1 Descriptive statistics:

Descriptive statistics of the overall pedagogical quality of lesson plans generated by Chat GPT and Copilot using TAG and CRISPE are presented in Table 03. The latter summarizes the averaged means for each experimental condition. Also, it demonstrates differences across models and frameworks.

Table 03 Pedagogical quality averaged means and differences

AI Model	CRISPE	TAG
<i>Chat GPT</i>	62.60	49.40
<i>Copilot</i>	54.80	48.80
Differences in means	7.8	0.6

The results indicate that the pedagogical quality of lesson plans varied depending on both the AI model and the framework. Specifically, lesson plans generated using CRISPE achieved substantially higher pedagogical quality scores than those generated by TAG, across both AI models. Particularly, under CRISPE framework, Chat GPT lesson plans obtained the highest mean score ($M=62.60$) compared to Copilot ($M= 54.80$). This implies that ChatGPT is more responsive to CRISPE than Copilot. In contrast, both AI models produced comparable moderate scores (Chat GPT= 49.40, Copilot= 48.80) when using TAG. This consistent moderate pedagogical quality scores across AI models may serve as preliminary evidence for possible constraints of the TAG framework.

Ultimately, these patterns demonstrate that structured frameworks like CRISPE can positively impact the quality of AI-generated lesson plans, particularly for ChatGPT. Indeed, this prompt framework guided the AI model to produce lesson plans that are all-inclusive, logically sequenced, well-organized and mostly aligned with objectives. Besides, they are relevant to the EFL context. In addition, they consistently target lower-order thinking skills

and some higher-order thinking skills such as analysis. The main issues associated with these plans are related to cognitive engagement and content reliability. Some of the suggested plans failed to target higher-order thinking skills: analysis, evaluation and creation. Also, their content was generic and inaccurate. Overall, effective implementation of these lesson plans necessitates thoughtful improvements.

Conversely, the results indicate that simpler prompt frameworks like TAG may generate mid-level outputs regardless of the model. Notably, TAG lesson plans meet some basic pedagogical requirements. However, they have multiple weaknesses regarding alignment, clarity, learner engagement and higher-order thinking skills. In other words, the AI-generated lesson plans under this condition are superficial and lack pedagogical depth. For instance, the suggested objectives, content and activities were partially aligned. Likewise, they support minimal cognitive engagement through the exclusive focus on recalling and basic understanding. Thus, these lesson plans require substantial revisions.

This descriptive comparison is inadequate to determine whether the observed differences and variances are statistically significant. Therefore, Two-way Analysis of Variance ANOVA (See Table 4) was conducted to examine the main effect of AI model, prompt framework, and their interaction on the lesson plans' pedagogical quality.

5.2 Two-Way Analysis of Variance

5.2.1. Levene's Test of Variance

Table 4 presents the results of Levene's test. Numerals refer to unequal variance of scores across the four cells, $F(3, 16) = 3.57, p = .038$. Thus, the assumption of homogeneity of variance is violated. However, it is acknowledged that Two-way ANOVA is robust to violation of homogeneity of variance when the cell sizes are equal (Field, 2013; Maxwell and Delaney, 2004; Glass, et al., 1972). On the grounds that the current study adopts a balanced factorial design (same size in each cell), the use of two-way ANOVA is statistically defensible and methodologically appropriate. Nevertheless, the results should be interpreted with caution.

Table 04 Levene's test

F	df1	df2	Sig
3.57	3	16	.038

5.2.2. Two-way ANOVA

Two-way ANOVA was conducted to examine the effects of AI models (ChatGPT, Copilot) and prompt frameworks (CRISPE, TAG) on lesson plans' pedagogical quality. As shown in table 5, the analysis of variance yielded a statistically significant main effect of the AI model $F(1, 16) = 33.60, p < .001$, partial $\eta^2 = .67$. The large effect size indicates that the AI model has a meaningful influence on the pedagogical quality of the lesson plan. However, it is not the only determinant of the overall quality. Other factors may moderate its impact.

Moreover, table 5 revealed a significant main effect of prompt framework $F(1, 16) = 175.54$, $p < .001$, with an extremely large effect size, partial $\eta^2 = .916$. This suggests a pronounced difference in pedagogical quality between lesson plans produced under CRISPE and those generated using TAG. Consequently, the prompt structure explains a substantial proportion of the variance within the statistical model and emerges as an influential factor in this study.

The interaction between AI model and prompt frameworks was also statistically significant $F(1, 16) = 64.28$, $p < .001$, with a large effect size partial $\eta^2 = .607$. Thus, the interaction explains a noticeable portion of variance in the quality of lesson plans. In other words, the effects of the AI model varied as a function of the prompt framework employed.

Accordingly, the findings of this study carry practical significance despite the modest sample size. The research hypotheses H_1 , H_2 , and H_3 are therefore supported. That is to say, the AI model, prompt framework and their interaction significantly influence the pedagogical quality of research methodology lesson plans.

Table 05 Tests of between subjects' effects

Dependent variable : Pedagogical quality						
Source	Type III sum of squares	Df	Mean square	F	Sig	Partial Eta Squared
Corrected Model	613,800 ^a	3	204,600	77,943	,000	,936
Intercept	58104,200	1	58104,200	22134,933	,000	,999
AI Model	88,200	1	88,200	33,600	,000	,677
Prompt Framework	460,800	1	460,800	175,543	,000	,916
AI model * Prompt Framework	64,800	1	64,800	24,686	,000	,607
Error	42,000	16	2,625			
Total	58760,000	20				
Corrected Total	655,800	19				

6. Discussion

This study sets out to evaluate the pedagogical quality and cognitive demand of AI-generated lesson plans. Especially, it examines the effects of AI models and prompt frameworks on the overall pedagogical quality of research methodology lesson plans. The results revealed that prompt framework has a significant impact on lessons' structural coherence and organization, clarity of objectives, pedagogical alignment, learner engagement, higher order thinking skills and adaptability. This supports the 2nd research hypothesis and aligns with the findings of Liu (2025), Dornburg and Davin (2025) who articulated the sensitivity of AI outputs to prompt engineering. Particularly, the CRISPE framework demonstrated a greater ability to scaffold instructional structure than the TAG framework. In other words, more structured frameworks appear to be more effective in enhancing the quality of AI output. This accentuates the role of the human educator as a guide and corroborates the

fact that AI cannot produce high-quality pedagogies without appropriate input. That is to say, AI can neither substitute nor supplant the human pedagogical expertise. Hence, AI should be used as an assistive tool to design effective pedagogical tasks not as time-saving shortcuts to finish highly demanding cognitive tasks.

Further, the significant main effect of AI model on the quality of the lesson plan provides evidence for the 1st research hypothesis and comports with Liu's conclusions. The findings indicate that ChatGPT, when properly guided, can produce "good" lesson plans. This model proved effective in generating adaptable lesson plans that are well structured and mostly aligned with learning objectives. However, its capacity to promote critical thinking skills remains limited. In contrast, Copilot's outputs failed to meet the expected standards. Indeed, multiple weaknesses relevant to alignment, learner engagement and higher-order thinking skills were identified. Ultimately, the variation in the quality of lesson plans across models are contingent on the underlying architecture of the model. It reflects the probabilistic and training patterns that govern the functioning of the AI model. This implies that AI models use different statistical patterns to produce generic responses rather than pedagogical expertise and nuanced judgement. Therefore, the results are likely to respect surface-level sequencing while lacking pedagogical alignment and cognitive engagement. This is why, some lesson plans appeared well structured, yet failed to target higher-order thinking skills. Indeed, the system that underpins AI models can reproduce patterns that emulate effective lesson plans. However, it cannot evaluate the soundness and rigor of the lesson plan. Consequently, human judgment and expertise are prerequisites for shaping the quality of AI instructional output.

In addition, the significant interaction effect between model and prompt framework on the lesson plans pedagogical quality supports the 3rd research hypothesis. The findings demonstrate that the effectiveness of prompt frameworks varies according to the AI model. This conforms with the study of Liu (2025) which suggests that some frameworks can promote comprehensive outputs across AI models.

Crucially, the moderate cognitive demand determined across all experimental conditions (4 cells) raises concerns and calls into question. Although AI-generated lesson plans are structurally coherent, they fall short in stimulating higher-order thinking skills like analysis, evaluation and creation, consistently and effectively. Therefore, the integration of AI into instructional design refers to an emergent practice that requires caution. Inevitably, this novel approach necessitates continuous critical evaluation of the accuracy, relevance, and depth of the AI-generated design for optimal results. Otherwise, cognitive rigor and learning outcomes will be compromised.

From an overall perspective, the findings articulate that AI models can support educators and teachers. However, they cannot function as autonomous pedagogical agents and instructional designers. Human expertise and prompt engineering skills are mandatory to maintain cognitive rigor, alignment and adaptability of AI outputs.

7. Limitations

While this study provides valuable insights into the effect of AI models and prompt frameworks on the pedagogical quality of lesson plans, the following limitations should be

acknowledged. First, the exploratory nature of the study precludes generalizations. Thus, the findings are contextually bounded. Particularly, the scope of the study is limited to research methodology, one of the most cognitively demanding subjects in the Algerian EFL curriculum. This implies that the findings may not apply uniformly to other subjects. Future research should replicate this study evaluating outputs on other modules like culture, grammar, written comprehension and literature. Second, although this study used only twenty lesson plans, a balanced design was established. This strengthened the robustness of ANOVA, even at a modest sample size. However, an increase in the number of lesson plans is likely to mitigate biases, entail robust interpretations and increase the generalizability of the findings. Third, only three raters were selected to evaluate 20 lesson plans. This can cause mental fatigue which may automatically affect evaluators' performance. To address this, future studies should consider recruiting more assessors. Fourth, this study uses two prompt frameworks and two AI models as the main source of data gathering. Relying merely on these models and frameworks can curtail the generalizability of findings. Consequently, future research should adopt other AI models and prompt frameworks. Lastly, this study did not implement any framework to identify hallucinated content across conditions. In this regard, it is recommended to use a comprehensive evaluation framework to assess the extent to which the AI content is valid and accurate.

8. Conclusion

This study explored the effectiveness of AI models and prompt structure on the overall pedagogical quality and cognitive demand of research methodology lesson plans. The findings demonstrate that both AI model and prompt structure had a statistically significant influence on the quality of the lesson plan. Particularly, lesson plans produced by ChatGPT under the CRISPE framework proved to be well-structured and pedagogically aligned.

Additionally, the findings revealed that the cognitive demand of AI-generated lesson plans across all conditions was restricted to the lower tiers of Bloom's taxonomy. In other words, these lesson plans failed to account for higher-order thinking skills consistently.

Therefore, AI lesson plans can be well-structured. Indeed, they can include the key pedagogical components. However, they may overlook the fundamental goal of education which is promoting critical thinking skills. Hence, adopting them without adjustments may potentially affect the intellectual rigor and learning outcomes.

A posteriori, human lesson plans can never be replaced by AI-generated lesson plans. Human pedagogical expertise is crucial to ensure that AI lesson plans comport with pedagogy and education standards. In other words, AI tools are assistive instruments, *per se*. They should be prudently implemented within a human guided, controlled, sound and ethical framework. This synergistic framework may entail excellence in education.

List of References:

1. Baytak, A. (2024). The content analysis of the lesson plans created by ChatGPT and Google Gemini. *Research in Social Sciences and Technology*, 9(1), 329–350. <https://doi.org/10.46303/ressat.2024.19>
2. Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75264–75278. <https://doi.org/10.1109/ACCESS.2020.2988510>
3. Dornburg, A., & Davin, K. J. (2025). ChatGPT in foreign language lesson plan creation: Trends, variability, and historical biases. *ReCALL*, 37(3), 332–347. <https://doi.org/10.1017/S0958344024000272>
4. Fan, H., Chen, G., Wang, X., & Peng, Z. (2024). LessonPlanner: Assisting novice teachers to prepare pedagogy-driven lesson plans with large language models. *arXiv*. <https://arxiv.org/abs/2408.01102>
5. Fang, Z., & Broussard, R. (2024). Generative AI in education: Opportunities, challenges, and implications for teaching and learning. *Educational Technology Research and Development*. <https://doi.org/10.1007/s11423-024-103XX-X>
6. Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE Publications
7. Iass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3), 237–288. <https://doi.org/10.3102/00346543042003237>
8. Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign
9. Kahraman, N., & Kılıcı, G. (2025). Evaluating the efficacy of AI-generated inquiry-based lesson plans in science. *Sakarya University Journal of Education*, 15(1), 40–53. <https://doi.org/10.19126/suje.1463067>
10. Karaman, M. R., & Göksu, İ. (2024). Are lesson plans created by ChatGPT more effective? An experimental study. *International Journal of Technology in Education*. <https://doi.org/10.46328/ijte.607>
11. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
12. Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). SAGE Publications.
13. Koryński, P., Woźniak, M., & Kamińska, D. (2023). Artificial intelligence in education: Applications and challenges. *Procedia Computer Science*, 217, 2210–2219. <https://doi.org/10.1016/j.procs.2023.01.203>
14. Lee, G.-G., & Zhai, X. (2024). *Using ChatGPT for science learning: A study on pre-service teachers' lesson planning*. *IEEE Transactions on Learning Technologies*. <https://doi.org/10.1109/TLT.2024.3401457>

15. Liu, X. (2025). *An evaluation of the pedagogical soundness and usability of AI-generated lesson plans across different models and prompt frameworks in high-school physics* (Preprint). arXiv. <https://arxiv.org/abs/2510.19866>
16. Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
17. Maxwell, S. E., & Delaney, H. D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (2nd ed.).
18. Powell, W., & Courchesne, S. (2024). *Opportunities and risks involved in using ChatGPT to create first grade science lesson plans*. PLOS ONE, 19(6), e0305337. <https://doi.org/10.1371/journal.pone.0305337>
19. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
20. Wang, F., Kinshuk, Chen, N.-S., & Spector, J. M. (2021). Artificial intelligence in education: Emerging technologies and future directions. *Educational Technology & Society*, 24(1), 1–8.
21. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., & Schmidt, D. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. arXiv. <https://arxiv.org/abs/2302.11382>