# INVESTING HOW AI GENERATED SUMMARIES AND PODCASTS INFLUENCE KNOWLEDGE RETENTION AND COMPREHNSION

By

Amrit K. Johal

A capstone project submitted for Graduation with University Honors

University Honors

University of California, Riverside

May 22, 2026

**APPROVED**

Dr. Rich Yueh
School of Business – Department of Information Systems

Dr. Begoña Echeverria, Howard H Hays, Jr. Chair
University Honors

## ABSTRACT

The rapid advancement of artificial intelligence (AI) has transformed how students access and use educational information. While AI-generated summaries and podcast-style explanations have become increasingly popular, less is known about how these formats influence immediate comprehension, delayed retention, engagement, and perceived understanding. This study compared AI-generated text summaries, AI-generated audio summaries, and human-written text summaries among undergraduate students enrolled in a business course at the University of California, Riverside. Participants were randomly assigned to one of the three summary-format conditions and completed an immediate comprehension assessment (T1), a delayed retention assessment approximately four hours later (T2), and post-learning measures of engagement, cognitive load, and confidence.

Results indicated no significant differences in immediate comprehension across summary formats. The predicted overall decline from T1 to T2 was not observed; instead, delayed scores were slightly higher than immediate scores, and item-level comparisons suggested that some T2 items were easier than their T1 counterparts. However, a significant time-by-format interaction showed that changes in quiz performance differed across conditions, with the AI-generated text condition showing a more favorable delayed-retention pattern than the human-written condition. Engagement and cognitive load did not significantly predict delayed retention after controlling for initial comprehension and summary format. Exploratory correlational analyses indicated that confidence and performance were positively related across all conditions, although delayed metacognitive accuracy appeared strongest in the human-written condition.

These findings suggest that AI-generated text may support short-term delayed recall under some conditions, but conclusions about overall retention should be interpreted cautiously because of quiz-version differences. As AI-assisted learning technologies continue to evolve, understanding how format shapes retention and self-monitoring may help educators and developers design more effective educational tools.

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my faculty mentor, Dr. Rich Yueh, for his continuous guidance, patience and support throughout my capstone project. His feedback and encouragement played a major role in shaping this research and my growth as a student researcher. I would also like to thank the University Honors department for their flexibility and support for helping students successfully complete meaningful research projects. Their encouragement made the process significantly manageable and rewarding.

I would also like to thank my family for their constant support throughout this journey. Their encouragement, understanding, and belief in me helped me stay motivated during the many stressful and uncertain moments that came with completing this project. I am especially thankful for their patience and reassurance throughout the entire process. Finally, I would like to thank everyone who contributed their time, guidance, and encouragement to me along the way.

**TABLE OF CONTENTS**

Abstract ..... 2

Acknowledgements ..... 4

Introduction ..... 6

Literature Review ..... 6

Hypotheses ..... 10

Method ..... 13

Participants..... 13

Materials..... 13

Procedure ..... 15

Results ..... 16

Discussion ..... 20

Limitations ..... 25

Conclusion ..... 27

References ..... 29

# INTRODUCTION

## *Literature Review*

Now more than ever, students are no longer learning content in the same ways they once did. Traditional methods such as textbooks and library-based research have gradually been supplemented, and in many cases replaced, by digital resources including online articles and video-based explanations. These developments significantly expanded access to information, allowing individuals with different learning styles and preferences to engage with content beyond static text (Moreno & Mayer, 2007). However, despite these shifts, learning remained fundamentally human-driven, with knowledge being created and communicated by human experts. In contrast, the current landscape represents a more dramatic transformation, where artificial intelligence tools such as ChatGPT and NotebookLM are increasingly used as primary sources of information. Research has shown that students are actively adopting AI tools for self-directed learning, leveraging their ability to generate explanations, answer questions, and produce study materials in real time (Ng et al., 2024; Ravšelj et al., 2025). Even standard web searches have evolved from lists of external sources to AI-generated summaries that deliver synthesized answers instantly. While these tools offer unprecedented speed and accessibility, it remains unclear whether they are capable of supporting meaningful understanding and long-term comprehension, as AI-generated responses may lack depth, contextual reasoning, and human nuance (Roll, 2016).

Due to the rapid pace of this technological shift, research has not kept up with how these tools influence actual learning outcomes. Existing studies have largely focused on AI adoption, user

perceptions, or general performance, rather than directly measuring comprehension and retention. Furthermore, there is a lack of experimental research comparing not only different AI models, but also the formats in which AI-generated information is delivered, such as text versus audio. This gap is particularly important given that format has been shown to influence how individuals process and retain information (Mayer, 2020). This study aims to address this gap by providing a contemporary evaluation of how modern students learn from AI-generated content, comparing its effectiveness to traditional human-written materials while also examining how different formats impact both learning outcomes and confidence in understanding.

Despite the clear advantages of AI-generated learning tools, there are significant limitations that raise concerns about their effectiveness in supporting meaningful learning. While these systems are designed to provide clear and efficient explanations, they often lack the depth, contextual understanding, and nuanced reasoning that human-generated content offers. Prior research has shown that AI-generated responses can appear structured and comprehensive, yet fail to fully capture the complexity of a topic, potentially leading to incomplete or oversimplified understanding (Roll, 2016). This is particularly concerning in educational contexts, where deep comprehension is essential for long-term knowledge retention.

Another major limitation lies in the potential for passive learning. Because AI tools deliver information in highly processed and easily digestible formats, users may engage less actively with the material. Cognitive learning research suggests that deeper processing and effortful engagement are critical for effective learning, and when information is presented too conveniently, it may reduce the level of cognitive effort required to understand it (Bjork & Bjork,

2011). As a result, students may develop a false sense of mastery, believing they understand the material when, in reality, their retention and ability to apply the information may be limited.

Additionally, while AI tools are often perceived as engaging and user-friendly, this perceived engagement does not necessarily translate into improved learning outcomes. Studies have found that students frequently report positive experiences when using AI-generated content, but also express concerns regarding accuracy, reliability, and depth of information (Ravšelj et al., 2025). This highlights a critical distinction between perceived learning and actual performance, suggesting that confidence in AI-generated explanations may not accurately reflect students' true level of understanding. Taken together, these limitations suggest that while AI tools may enhance accessibility and engagement, they may also introduce risks related to shallow processing, overconfidence, and reduced retention. These concerns underscore the importance of evaluating not only how AI-generated content is used, but also how effectively it supports meaningful learning outcomes.

Beyond the general advantages and limitations of AI-generated content, the format in which information is presented plays a critical role in how effectively it is processed and retained. Cognitive theories of multimedia learning suggest that different modes of information delivery, such as text and audio, engage distinct cognitive processes that can influence comprehension and memory (Mayer, 2020). For example, text-based learning allows individuals to control the pace at which they consume information, revisit specific sections, and actively engage with the material through rereading and reflection. This level of control can support deeper processing and more detailed understanding. In contrast, audio-based learning relies on sequential

processing, where information is delivered continuously and cannot be easily revisited without repetition. While this format may enhance engagement and leverage natural listening and storytelling mechanisms, it may also increase cognitive load if learners are unable to process information at the pace it is presented (Moreno & Mayer, 2007).

Prior research has explored the differences between reading and listening in traditional learning environments, often highlighting trade-offs between comprehension, engagement, and retention. However, these findings have largely been examined independently of AI-generated content. As AI tools increasingly produce both written summaries and podcast-style audio explanations, it becomes important to understand whether these established differences in learning formats persist, diminish, or change when applied to AI-generated materials. While text-based AI tools such as OpenAI's ChatGPT are commonly used for quick reference and structured study, audio-based tools like Google's NotebookLM are gaining popularity for their convenience and ability to support passive learning during activities such as commuting or multitasking. Despite growing research on AI in education and established theories on learning formats, there remains a clear gap in understanding how these two areas intersect. Existing studies have primarily focused on AI adoption, user perceptions, or isolated learning outcomes, while research on text versus audio learning has largely been conducted outside the context of AI-generated content.

As a result, little is known about how AI-generated text and audio summaries compare in their ability to support both immediate comprehension and long-term retention. Furthermore, prior research has not sufficiently examined how these formats influence perceived understanding, particularly in relation to actual performance outcomes. To address these gaps, this study

conducts a controlled experimental comparison of three learning conditions: AI-generated text summaries, AI-generated audio summaries, and human-written text summaries. By measuring both immediate comprehension and delayed retention, this study provides a more comprehensive evaluation of learning effectiveness across formats. The study introduces a first set of questions (T1) and then an additional set of questions taken 4-8 hours after the original (T2). In addition, the study incorporates measures of engagement, mental effort, and confidence to better understand how users perceive their learning experiences in both sets. The goal of this research is to determine not only which format leads to stronger learning outcomes, but also whether AI-generated content influences the relationship between perceived and actual understanding. As AI continues to play an increasingly central role in education, these findings can provide valuable insights for students, educators, and developers seeking to design more effective and reliable learning tools.

### *Hypotheses*

In this study, there are four sets of hypotheses (H1, H2, H3, H4), with two complementary analyses.

**H1: Immediate Comprehension (T1 Quiz Scores):** There will be a significant difference in quiz performance (T1) across all three summary formats (AI-generated text summary, AI-generated audio summary, and human-written text summary).

**H2: Delayed Retention (T2 Quiz Scores):** Participants across all conditions will perform worse on the delayed retention quiz (T2) compared to the immediate comprehension quiz (T1).

**H3:** The amount of decline from T1 to T2 will vary by summary format. Participants in the AI-generated text condition are expected to show a smaller decline than participants in the AI-generated audio condition.

**H4:** Higher engagement, measured with the UES-SF, and lower cognitive load / mental effort are expected to predict better delayed retention performance (T2), even after controlling for T1 quiz scores and summary format.

In regards to H1, this hypothesis predicts that there will be significant differences in immediate comprehension performance across the three summary formats. Prior multimedia learning research suggests that the format in which information is delivered can influence how effectively learners process and retain information (Mayer, 2020). Text-based learning formats allow individuals to control the pace at which they consume information, revisit sections of content, and engage in deeper cognitive processing. In contrast, audio-based learning relies on continuous sequential processing, making it more difficult for learners to pause, review, and reflect on information during consumption. Because of these differences in information processing, it is expected that summary format will significantly influence participants' T1 quiz performance.

H2 predicts that participants across all conditions will demonstrate lower performance on the delayed retention quiz (T2) compared to the immediate comprehension quiz (T1). This prediction is based on established learning and memory research demonstrating that information retention naturally declines over time without reinforcement or repeated retrieval. While all groups are expected to experience some level of decline between T1 and T2, this study aims to evaluate whether delayed retention differs depending on how information was initially presented.

H3 examines whether the magnitude of retention change differs across summary formats. Specifically, participants in the AI-generated text condition are expected to demonstrate a smaller decline in performance between T1 and T2 than participants in the AI-generated audio condition. This prediction is grounded in multimedia learning theory, which suggests that text-based formats may support stronger delayed retention by allowing learners to revisit and process information more actively, whereas audio-based formats may prioritize engagement and convenience over deeper cognitive processing. Because this study measures both immediate comprehension and delayed retention, differences in learning effectiveness between formats are expected to become more apparent over time.

H4 examines whether engagement and cognitive load predict delayed retention performance after accounting for initial comprehension and summary format. Specifically, participants who report higher engagement and lower cognitive load are expected to demonstrate stronger T2 quiz performance. Prior research on AI-assisted learning has suggested that students often perceive AI-generated educational materials as accessible, engaging, and user-friendly, particularly when information is presented in concise and digestible ways (Ng et al., 2024). At the same time, cognitive load theory suggests that excessive mental effort can reduce learning efficiency and impair retention. As a result, participants who experience stronger engagement while also perceiving lower cognitive strain are expected to demonstrate more effective learning outcomes, even after controlling for initial quiz performance and assigned summary format.

## METHOD

### *Participants*

Participants were undergraduate students recruited from a business course at the University of California, Riverside. Students were offered the opportunity to participate in the study as one option for completing an extra credit assignment for the course. Students were eligible to receive partial extra credit for completing the immediate assessment (T1) and additional extra credit for completing the delayed follow-up assessment (T2). Students who did not wish to participate in the study were provided with an alternative assignment of equal credit value.

Before participating, students read an informed consent form that described the study's goals, methods, voluntary nature, confidentiality safeguards, and the researcher's contact details. The course instructor did not have access to participant responses or study data in order to protect anonymity and minimize potential bias. Teaching assistants were only informed of completion status of students for the purposes of rewarding extra credit.

### *Materials*

The study materials consisted of pre-survey measures, learning materials, comprehension assessments, and post-learning questionnaires administered electronically through Qualtrics.

The pre-survey collected demographic and background information including age, gender, prior familiarity with the study topic, preferred learning format, self-rated working memory, and prior use of AI-assisted learning tools. These variables were included to better understand participant characteristics and potential differences in learning outcomes.

Participants were assigned to one of three learning conditions: AI-generated text summaries, AI-generated audio summaries, or human-written text summaries. The AI-generated text materials were created using ChatGPT, while the AI-generated audio materials were generated using NotebookLM's podcast-style audio summary feature. The human-written condition consisted of a traditionally written summary created from expert-written educational content. Across all conditions, materials were standardized as closely as possible in length, informational content, and overall topic coverage to minimize differences unrelated to format.

Following exposure to the learning materials, participants completed a T1 comprehension quiz consisting of multiple-choice questions designed to measure immediate understanding of the content. Approximately four hours later, participants completed a second quiz (T2) measuring delayed retention of the same information. The four-hour delay was selected to evaluate short-term retention decline while also improving follow-up completion rates relative to longer delay intervals commonly used in educational research.

Participants also completed post-learning survey measures assessing engagement, cognitive load, distraction, perceived quiz performance, confidence in understanding, and confidence in recall. Additional manipulation check items measured perceived clarity, detail, and perceived AI authorship of the learning materials. Engagement was measured using the User Engagement Scale Short Form (UES-SF), with negatively worded items reverse-coded prior to reliability analysis (O'Brien et al., 2018).

### *Procedure*

Participants first accessed the study through a Qualtrics survey link and reviewed an informed consent form describing the study procedures, voluntary participation, confidentiality protections, and extra credit requirements. After providing consent, participants completed a pre-survey measuring demographic information, prior familiarity with the study topic, learning-format preferences, self-rated working memory, and prior AI-tool usage.

Participants were then randomly assigned to one of three experimental conditions. Participants in the AI-generated text condition read a text-based summary generated using ChatGPT. Participants in the AI-generated audio condition listened to a podcast-style audio explanation generated through NotebookLM. Participants in the human-written text condition read a traditional expert-written summary covering the same topic. Participants were instructed to carefully review the material because they would complete quizzes based on the information presented.

Immediately after interacting with the learning material, participants completed the T1 comprehension quiz followed by post-learning survey measures assessing engagement, cognitive load, distraction, perceived quiz performance, and confidence in understanding. Approximately four hours after completing the initial portion of the study, participants received an automated Qualtrics email containing a link to the T2 follow-up assessment. During the delayed assessment, participants completed a second quiz measuring retention of the previously learned material, along with additional questions related to confidence in recall and perceived memory of the

content. Participants were eligible to receive partial extra credit for completing T1 and full extra credit for completing both T1 and T2.

## RESULTS

A total of 101 participants completed both the immediate comprehension assessment (T1) and delayed retention assessment (T2) across the three experimental conditions. The final analytic sample included participants who completed both the immediate assessment and the delayed follow-up assessment. Participants were assigned to one of three summary formats: human-written text (n = 32), AI-generated audio (n = 28), and AI-generated text (n = 41).

To test H1, a one-way between-subjects ANOVA was conducted to examine whether immediate comprehension performance differed across summary formats. Results indicated that there was not a statistically significant difference in T1 quiz scores between the human-written text condition (M = 7.31, SD = 1.91), AI-generated audio condition (M = 7.21, SD = 2.30), and AI-generated text condition (M = 6.49, SD = 1.83), $F(2, 98) = 1.88$, $p = .159$, $\eta^2 = .037$. Post hoc Tukey comparisons also revealed no significant pairwise differences between conditions. Accordingly, H1 was not supported. These findings suggest that summary format did not significantly influence immediate comprehension performance immediately after exposure to the learning materials.

To evaluate H2, paired-samples t-tests were conducted comparing immediate comprehension scores (T1) and delayed retention scores (T2). Initial analyses indicated that the difference between T1 and T2 scores approached statistical significance, $t(100) = -1.92$, $p = .058$. However,

assumption testing revealed significant deviations from normality based on the Shapiro-Wilk test, $W = .945$, $p < .001$. As a result, a Wilcoxon signed-rank test was conducted as a nonparametric follow-up analysis. The Wilcoxon test revealed a statistically significant difference between T1 and T2 scores, $z = -2.25$, $p = .022$. Interestingly, overall delayed retention scores ($M = 7.34$, $SD = 2.24$) were slightly higher than immediate comprehension scores ($M = 6.95$, $SD = 2.01$), suggesting that retention performance did not decline overall as originally predicted. Accordingly, H2 was not supported.

Informal item-level checks suggested that some T2 items may have been easier than their T1 counterparts. For example, external item-level calculations indicated that Question 3 and Question 9 had noticeably higher correct-response rates at T2 than at T1. This pattern suggests that the unexpected improvement from T1 to T2 may partly reflect quiz-version differences rather than true retention gains.

Although overall retention decline was not observed across all participants, differences emerged when examining retention change across summary formats. To test H3, a repeated-measures ANOVA was conducted examining the interaction between time (T1 versus T2) and summary format. Results revealed a statistically significant interaction effect between time and summary format, $F(2, 98) = 4.43$, $p = .014$, $\omega^2 = .012$. This finding indicates that changes in quiz performance over time differed significantly depending on the assigned learning format.

Further analysis examining retention change scores revealed that participants in the human-written text condition demonstrated the largest decline in performance between T1 and T2 ($M = -0.38$, $SD = 2.08$), while participants in the AI-generated audio condition showed a slight

increase in scores over time (M = 0.36, SD = 2.08). Participants in the AI-generated text condition demonstrated the strongest improvement in retention performance over time (M = 1.00, SD = 1.78). A one-way ANOVA examining retention change across conditions was statistically significant, $F(2, 98) = 4.43$, $p = .014$, $\omega^2 = .064$. Post hoc Tukey comparisons revealed that the AI-generated text condition differed significantly from the human-written text condition ($p = .010$), while differences between the AI-generated audio and AI-generated text conditions were not statistically significant. These findings indicate that changes in quiz performance over time differed significantly by summary format. However, the specific predicted contrast between the AI-generated text and AI-generated audio conditions was not statistically significant in post hoc testing. Instead, the clearest difference emerged between the AI-generated text and human-written text conditions. Accordingly, H3 received partial support.

To evaluate H4, hierarchical linear regression analyses were conducted examining whether engagement and cognitive load predicted delayed retention performance after controlling for initial comprehension and summary format. In the first regression model, T1 quiz scores significantly predicted T2 quiz scores, $\beta = .551$, $p < .001$, explaining approximately 30.4% of the variance in delayed retention performance. After adding summary format variables in the second step, the model explained approximately 34.6% of the variance, with the AI-generated text condition significantly predicting higher T2 quiz scores relative to the human-written condition, $p = .015$. Accordingly, H4 was not supported.

In the final regression model, engagement (UES-SF total score) and cognitive load (mental effort scale) were added as predictors.The scale demonstrated relatively low internal consistency

reliability (Cronbach's $\alpha = .440$), even after reverse coding negatively worded items. Due to the low reliability estimate, findings involving engagement measures were interpreted cautiously throughout the analysis. However, neither engagement ($p = .539$) nor cognitive load ($p = .235$) significantly predicted delayed retention performance after accounting for T1 quiz scores and summary format. Additionally, the increase in explained variance after adding engagement and cognitive load variables was not statistically significant, $\Delta R^2 = .010$, $p = .472$. These findings suggest that initial comprehension performance was a substantially stronger predictor of delayed retention than subjective engagement or perceived mental effort.

Complementary analyses indicated no significant differences across formats in either engagement, $F(2, 98) = 0.70$, $p = .500$, or mental effort, $F(2, 98) = 1.34$, $p = .266$. Exploratory analyses also examined the relationship between participants' confidence and their quiz performance. Overall, confidence was positively associated with performance, suggesting that participants had some ability to judge their own understanding. However, these relationships varied somewhat across summary formats. In the human-written condition, T1 confidence was strongly correlated with both T1 quiz performance and T2 quiz performance. Similar positive relationships were observed in the AI-generated audio and AI-generated text conditions, although the associations with later T2 performance were somewhat weaker. These findings suggest that participants' confidence generally aligned with their actual performance, but confidence may have been a less consistent indicator of delayed retention in the AI-generated conditions. Because these analyses were exploratory and correlational, they should be interpreted cautiously. Future research should examine metacognitive accuracy more directly by comparing participants' confidence ratings with performance across equivalent quiz forms and longer retention intervals.

## DISCUSSION

The purpose of this study was to examine how AI-generated text summaries, AI-generated audio summaries, and human-written text summaries influence immediate comprehension, delayed retention, engagement, cognitive load, and perceived understanding. As artificial intelligence tools become increasingly integrated into educational environments, it is important to understand not only whether these tools improve learning outcomes, but also how different AI-generated formats shape delayed recall and learners' judgments about their own understanding. Overall, immediate comprehension did not differ significantly across formats. The predicted overall decline from T1 to T2 was not observed, and item-level analyses suggested that differences in quiz difficulty likely contributed to the unexpected T2 pattern. At the same time, retention change differed across formats, with the AI-generated text condition showing a more favorable delayed-retention pattern than the human-written condition. Engagement and cognitive load did not significantly predict delayed retention after controlling for initial comprehension and summary format. Exploratory correlational analyses also suggested that confidence and performance were positively related across all conditions, although delayed metacognitive accuracy appeared strongest in the human-written condition.

The first hypothesis predicted that there would be significant differences in immediate comprehension performance across summary formats. However, H1 was not supported, as participants across the AI-generated text, AI-generated audio, and human-written text conditions performed similarly on the T1 quiz. These findings suggest that all three formats were similarly effective in supporting immediate understanding of the material immediately after exposure. One

explanation for this outcome may be that all participants completed the comprehension assessment shortly after interacting with the learning materials, limiting the amount of time for forgetting or deeper processing differences to emerge. Additionally, the summaries across conditions were intentionally standardized in content, length, and structure, which may have reduced variability in immediate learning performance. These findings align with prior multimedia learning research suggesting that when informational content is equivalent, immediate comprehension outcomes across formats may appear relatively similar in short-term testing environments (Mayer, 2020).

Although H1 was not supported, delayed retention findings provided a more nuanced understanding of how summary format influences learning over time. H2 predicted that participants would demonstrate lower performance on the delayed retention assessment compared to the immediate comprehension assessment. This hypothesis was not supported. While the paired-samples t-test approached significance ($p = .058$), the nonparametric Wilcoxon signed-rank test revealed a statistically significant difference between T1 and T2 scores. However, overall T2 scores were slightly higher than T1 scores rather than lower as originally predicted. Item-level comparisons suggest that this unexpected pattern likely reflects, at least in part, differences in difficulty between the T1 and T2 quiz versions rather than genuine improvement in retention. In particular, Question 3 and Question 9 showed substantial increases in correct responding at T2, indicating that the reworded versions of these items were easier than their T1 counterparts. It is also possible that completing the T1 quiz produced a retrieval-practice benefit, which prior memory research has shown can improve later recall performance through

the testing effect (Roediger & Karpicke, 2006). However, the item-level evidence suggests that quiz-version nonequivalence is the more immediate concern.

The most important findings of the study emerged through H3, which examined whether retention change differed across summary formats. H3 received partial support. The repeated-measures ANOVA showed a significant time × summary format interaction, indicating that changes in performance over time varied by condition. However, the specific predicted contrast between the AI-generated text and AI-generated audio conditions was not statistically significant; instead, the clearest difference emerged between the AI-generated text and human-written conditions. These findings suggest that differences between learning formats may become more apparent over time rather than immediately after exposure.

The superior retention performance observed in the AI-generated text condition may be explained through cognitive processing and multimedia learning theories. Unlike audio-based learning, text-based formats allow learners to revisit information, reread confusing sections, and process content at their own pace. This level of control may support deeper encoding and stronger memory consolidation over time. Additionally, AI-generated summaries are often structured in highly simplified and organized formats that reduce informational complexity while maintaining clarity. This may have made it easier for participants to mentally organize and retrieve information during the delayed assessment. These findings partially support prior research suggesting that concise and digestible AI-generated educational materials can improve accessibility and reduce barriers to comprehension (Ng et al., 2024). At the same time, these

results should be interpreted cautiously because item-level analyses suggest that the T1 and T2 quiz versions were not perfectly equivalent in difficulty.

The AI-generated audio findings were also notable. Although participants in the audio condition did not perform significantly worse than other groups on immediate comprehension, their retention performance did not improve to the same extent as the AI-generated text group. This may reflect differences in how learners process spoken information compared to written information. Audio-based learning relies on sequential processing, meaning learners cannot easily revisit earlier information without replaying the material. While podcast-style formats may feel more engaging and conversational, they may reduce opportunities for active review and reflection during learning. These findings support multimedia learning theories suggesting that convenience and engagement do not always translate into stronger retention outcomes.

H4 examined whether engagement and cognitive load predicted delayed retention performance after controlling for initial comprehension and summary format. Contrary to expectations, neither engagement nor cognitive load significantly predicted T2 performance. Instead, initial comprehension performance emerged as the strongest predictor of delayed retention. These findings suggest that subjective perceptions of engagement and mental effort may be less important for retention than the quality of initial encoding and understanding. Although participants may feel highly engaged while consuming educational content, engagement alone may not guarantee stronger memory retention or deeper learning. This distinction is particularly important in AI-assisted educational environments, where highly interactive or entertaining formats may create the impression of effective learning without necessarily improving long-term

retention. This interpretation should be made cautiously, however, because the UES-SF demonstrated poor internal consistency in this sample.

The findings related to perceived understanding and confidence were also informative. Rather than showing broad overconfidence in the AI conditions, the clearest pattern was a difference in metacognitive accuracy over time. Immediately after learning, participants in all three conditions showed strong positive relationships between confidence and quiz performance, suggesting that they could judge their immediate comprehension reasonably well. At the delayed test, however, metacognitive accuracy remained strongest in the human-written condition, whereas confidence-performance relationships were weaker in the AI-generated conditions, particularly for recall confidence. This pattern suggests that AI-generated materials may not simply inflate confidence; instead, they may make it harder for learners to monitor what they have actually retained over time. If students are less accurate in judging their delayed knowledge, they may be less likely to identify gaps in understanding and seek additional review. Future research should examine whether this pattern persists over longer retention intervals and with other forms of AI-generated content.

These findings have important implications for students, educators, and developers of AI learning technologies. For students, the results suggest that AI-generated text summaries may serve as effective study tools for reinforcing retention and supporting delayed recall. However, students should remain cautious about assuming that engaging or convenient formats automatically lead to stronger learning outcomes. For educators, the findings highlight the importance of critically evaluating how AI tools are integrated into learning environments rather

than treating all AI-generated formats as equally effective. While podcast-style and conversational AI formats may increase accessibility and convenience, text-based AI summaries may provide stronger support for long-term retention and review-based studying.

The findings also have implications for developers of AI educational tools. Many modern AI platforms prioritize conversational interaction, entertainment, and ease of consumption, particularly through audio and podcast-style features. However, this study suggests that features promoting active review, self-paced learning, and information revisiting may be more beneficial for long-term learning outcomes. Developers should consider designing AI-assisted educational tools that balance engagement with opportunities for deeper cognitive processing and retrieval practice. As AI continues to become more integrated into educational systems, understanding how format influences learning effectiveness will become increasingly important for optimizing student outcomes.

### *Limitations*

Several limitations should be considered when interpreting the findings of this study. First, the participant sample consisted entirely of undergraduate students recruited from a single business course at a large public university in the southwestern United States. As a result, the findings may not generalize to broader populations, including students from different academic disciplines, universities, age groups, or educational backgrounds. Future research should examine whether similar patterns emerge across more diverse and representative samples.

Second, the study utilized a relatively short four-hour retention interval between the immediate comprehension assessment (T1) and delayed retention assessment (T2). While this delay period allowed the study to examine short-term retention changes while maintaining participant completion rates, it does not fully capture long-term learning outcomes. Future studies should investigate whether the observed differences between summary formats persist across longer retention intervals such as several days or weeks.

One limitation concerns quiz equivalence across timepoints. Although the T1 and T2 quizzes were designed to assess the same content, item-level analyses indicated that some T2 items appeared easier than their T1 counterparts. This complicates interpretation of the overall T1–T2 comparison because differences in performance may reflect changes in item difficulty rather than changes in retention alone. Future studies should devote greater attention to establishing equivalence between alternate quiz forms before comparing performance across time.

Another limitation involves the experimental nature of the study setting. Participants interacted with the learning materials in a controlled online environment and completed quizzes shortly afterward, which may not fully reflect how students naturally use AI tools in real-world educational settings. In practice, students often use AI-generated summaries alongside additional resources, repeated review sessions, note-taking strategies, or multitasking behaviors that were not captured in this study. Future research should examine AI-assisted learning behaviors in more naturalistic academic contexts. The study also relied heavily on self-reported measures of engagement, cognitive load, confidence, distraction, and perceived understanding. Self-report measures are inherently subjective and may not always accurately reflect actual cognitive

processing or learning experiences. Additionally, the User Engagement Scale Short Form (UES-SF) demonstrated relatively low internal consistency reliability within this sample, limiting the strength of conclusions involving engagement measures. Future research should incorporate additional objective measures of engagement and cognitive effort, such as behavioral tracking, eye-tracking technology, or physiological indicators of attention.

Finally, although efforts were made to standardize content across conditions, differences between AI-generated and human-written materials may still have influenced learning outcomes in ways beyond format alone. AI-generated summaries may differ from human-written summaries in tone, organization, sentence structure, or simplification style, making it difficult to isolate format as the only contributing factor. Future studies should continue refining content standardization procedures when comparing AI-generated and traditional educational materials.

## CONCLUSION

This study examined how AI-generated text summaries, AI-generated audio summaries, and human-written text summaries influence immediate comprehension, delayed retention, engagement, cognitive load, and perceived understanding. As artificial intelligence tools become increasingly integrated into educational environments, understanding how these formats affect learning outcomes is important for both students and educators. While immediate comprehension did not significantly differ across summary formats, change over time did vary by format, with the AI-generated text condition showing a more favorable delayed-retention pattern than the human-written condition.

At the same time, the expected overall decline from T1 to T2 was not observed, and item-level analyses suggest that differences between the two quiz versions likely contributed to the unexpected overall time pattern. Engagement and cognitive load did not significantly predict delayed retention after accounting for initial comprehension and summary format. Exploratory correlational analyses suggested moderate positive relationships between confidence and performance across all conditions, although delayed metacognitive accuracy appeared strongest in the human-written condition.

Overall, these findings highlight the importance of evaluating not only whether AI-generated educational tools are effective, but also how different formats shape delayed recall and learners' ability to monitor their own understanding. Future research should examine longer retention intervals, more diverse student populations, and more carefully matched quiz forms to better understand how artificial intelligence can most effectively support meaningful learning outcomes.

## REFERENCES


Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (2nd ed., pp. 56–64). Worth Publishers.

Mayer, R. E. (2020). Multimedia learning (3rd ed.). Cambridge University Press.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. Educational Psychology Review, 19(3), 309–326.

Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. British Journal of Educational Technology, 55(4), 1328–1353.

O'Brien, H. L., & Toms, E. G. (2018). The development and evaluation of a survey to measure user engagement. Journal of the American Society for Information Science and Technology.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. Journal of Educational Psychology, 84(4), 429–434.

Ravšelj, D., Keržič, D., & Tomaževič, N. (2025). Higher education students' perceptions of ChatGPT: A global study of early reactions. PLoS One, 20(2).

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roll, I. (2016). Evolution and revolution in artificial intelligence in education. International Journal of Artificial Intelligence in Education, 26(2), 582–599.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Taskiran, N. (2023). Effect of artificial intelligence course in nursing on students' medical artificial intelligence readiness: A comparative quasi-experimental study. Nurse Educator, 48(5), E147–E152.