Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
April 23, 2026
A Regional Conference of the Society for Industrial and Systems Engineering

Assessing the Impact of AI-Based Meeting Summarization on Information Relevance and Engagement

James Spoerl, Daniel Choi, Jeremy Locklear, Henry Engler, Zachary Leith, and Donald Koban

Department of Systems Engineering, United States Military Academy, West Point, New York 10996

Corresponding author's Email: donald.koban@westpoint.edu

Author Note: This work was supported by the Special Operations Command (SOCOM) Ignite Program. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of War.

Abstract: Modern professional environments generate high volumes of communication, leaving limited time to process information. Although Artificial (AI)-based summarization tools are widely adopted, there is limited controlled evidence quantifying their incremental impact on user engagement. This study tests whether AI-generated summaries influence perceived relevance and listening intentions relative to human-written summaries. In a randomized within-participant experiment (N = 304), cadets evaluated both summary types across multiple topic domains. AI-generated summaries increased perceived relevance and listening intentions in some topic domains and showed no measurable disadvantage in others. These findings suggest that AI-based summarization can support engagement in institutional communication settings, but that effectiveness depends on content alignment. Future research should test these effects in more diverse contexts and evaluate whether optimized prompting or retrieval-augmented generation (RAG) further enhances engagement.

Keywords: Text Summarization, Large Language Models, Perceived Relevance, Behavioral Intentions

1. Introduction

Modern leaders are not constrained by access to information, but by time and attention. Professionals across many fields manage heavy workloads, overlapping responsibilities, and constant communication, often without additional time to process it all. A Harvard Business Review study found that 38% of employees report receiving an excessive volume of communications, and 27% say they feel overloaded by the information they must process (Klein et al., 2023). In fast-paced environments, frequent meetings generate large amounts of information that must be captured, interpreted, and referenced later to guide decisions. When summaries are created under time pressure, manual workflows can vary in quality, consistency, and timeliness. For teams operating under constant information demands, the ability to quickly distill key insights is not simply convenient; it can create a decision advantage. Artificial Intelligence (AI)-based summarization offers a scalable way to standardize information capture and reduce the burden of managing limited time.

As AI-generated summaries have become more fluent and coherent, the central question has shifted from whether models can generate readable text to how their impact should be evaluated. Traditional evaluation relies on lexical overlap metrics such as Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), which measure n-gram similarity between system outputs and reference texts. Although scalable, these metrics define relevance in terms of surface correspondence and often show limited alignment with human judgment. Research in trust in automation similarly emphasizes that system performance alone does not guarantee effective use; value depends on whether outputs support calibrated reliance and align with user expectations (Lee & See, 2004).

Subsequent work sought closer alignment with meaning rather than surface correspondence. Semantic similarity measures such as BERTScore (Zhang et al., 2019) and innovations such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) aimed to ground summaries in source material and improve faithfulness. More recent efforts, such as AUTOCALIBRATE (Liu et al. 2023), optimize large language model prompts to increase agreement between automated evaluation and expert human ratings. Applied systems have also emphasized contextual enrichment and personalization, such as the multi-source RAG-based meeting summarization framework developed by Kirstein et al. (2024). With recent advances in large language models improving fluency and coherence (Laskar et al., 2023), practical trade-offs now center more on cost, speed, and privacy than baseline readability.

Despite these technical advances, most summarization research continues to evaluate relevance at the model level. In contrast, foundational work in information retrieval defines relevance as user-dependent and context-sensitive (Peikos & Pasi, 2024; Samimi & Ravana, 2014). Consistent with this perspective, operational studies of Large Language Model (LLM)-enabled

workflows show that value depends not only on accuracy but on whether users interpret and trust the outputs (Belveal et al., 2025). Building on these insights, the present study moves beyond algorithmic scoring and evaluates perceived relevance and engagement intentions in an applied institutional setting. Rather than asking whether a summary matches a reference text, we ask whether it increases perceived relevance and intentions to engage further, which reflect early signals that users are willing to rely on the output.

In collaboration with a Special Operations Command (SOCOM) unit seeking to reduce information overload in high-volume meeting environments, we used West Point Press podcasts as a controlled test case to examine whether AI-generated summaries influence perceived relevance and engagement intentions. In a within-participant experiment (N = 304), cadets reviewed both AI-generated and human-written summaries and completed survey measures assessing perceived relevance and their intention to engage further with the full content. We focus on perceived relevance because people direct limited attention toward information they believe is worth their time. Relevance judgments often serve as a first step toward deeper engagement and informed decision-making. We hypothesized that AI-generated summaries would increase perceived relevance and strengthen engagement intentions compared to human-written summaries.

To our knowledge, no prior controlled experiments have examined how AI-generated summaries affect user-level engagement outcomes in a military professional education setting. By shifting the focus from algorithmic accuracy to user behavior, this study provides empirical evidence to inform decisions about adopting AI-based summarization as a scalable tool for managing information in time-constrained institutional environments.

2. Methodology

2.1 Design

To evaluate the effect of automated summarization on user engagement, we conducted a randomized, within-participant experiment comparing AI-generated summaries to human-written summaries. The primary outcomes were perceived relevance of the summary and intention to listen to the associated podcast episode. Participants were randomized to one of four counterbalanced presentation paths. Each path included four summaries drawn from distinct topical domains: Leadership, Data, Technology, and Strategy (See Table 1), with condition assignment and order rotated to mitigate order and topic effects. An instructional attention check was embedded midway through the sequence. This counterbalanced design allowed each participant to evaluate both treatment and control summaries while minimizing sequencing artifacts.

Table 1. Podcast episodes used as experimental stimuli, organized by topical domain.

Topic	Episode Title
Leadership	*Winning Culture with Coach K: Leadership Lessons on Building Teams and Champions*
	Retaining Good Talent Through Inspiring Leadership: A Discussion with Colonel Everett Spain
Data	*Decoding Data Literacy: Perspectives from Colonel Nicholas Clark*
	Bridging Human and Machine through Data Science with Colonel David Beskow
Technology	*Exploring Space and Modern Warfare with Lt. Col. Will Koch*
	Leadership & Technology on the Battlefield with General (Ret.) Scott Miller
Strategy	*Global Security: NATO Secretary General Mark Rutte at West Point*
	Building Cross-Cultural Competence Through Language & Cultural Immersion with Col. John Baskerville

2.2 Materials

The study used summaries from eight episodes of the United States Military Academy Dean's Podcast. For each episode, two versions were created: (1) the original human-written podcast description (control condition) and (2) an AI-generated summary produced using Microsoft Copilot (treatment condition). Example summaries from both conditions are presented in Table 2. AI summaries were generated using a standardized prompt designed to ensure structural consistency. Each summary followed a fixed three-part format: (1) a ~50-word introduction highlighting leadership relevance, (2) exactly three action-oriented insights structured around a threat–requirement–payoff sequence, and (3) a single-sentence motivational takeaway. Word limits, tone, and rhetorical style were constrained to maintain consistency across episodes. Copilot was instructed to generate multiple internal drafts and return one final version. All content had to be drawn strictly from the original transcript, and no manual editing was applied after generation. Participants were not told whether a summary was AI- or human-generated. Each participant evaluated four summaries in total - two AI-generated and two human-written - presented in randomized order based on their assigned path.

Table 2. Example human-written and AI-generated (Microsoft Copilot) summaries for the Data Literacy podcast episode.

Human Summary	CoPilot Summary
Join us for another episode of *Inside West Point: Ideas That Impact.* In this episode, Colonel Nicholas Clark and West Point Dean Brigadier General Shane Reeves explore the pivotal role of data-driven decisions in the military and. Clark sheds light on the inception of the Center for Data Analysis and Statistics, his creation of a data literacy program for the Army, and the surging significance of data and analytics. Clark underscores the imperative for officers to possess data literacy and unveils his research award focused on engaging undergraduates in data science. Tune in to discover how to enhance your data literacy and gain insights into the fulfilling realm of teaching at West Point.	Afghanistan's withdrawal exposed a silent threat: leaders drowning in messy, unstructured data. Colonel Nicholas Clark warns that without "data-informed" judgment, commanders risk paralysis at the point of decision. Officers who master data literacy—asking the right questions, shaping raw inputs, and communicating insights—will weaponize information for speed, clarity, and mission success. Three Key Insights: • Look for chaotic "data deserts" and overwhelming raw feeds that obscure operational decisions. • Strive for disciplined data literacy—ask sharper questions, format inputs, and enforce analytics in briefs. • Gain faster decisions and readiness by turning sensors into solutions at the edge of combat. Listen to Colonel Clark explain why data literacy is now a warfighting imperative for every Army leader.

2.3 Participants and Procedure

Participants were cadets enrolled in required Behavioral Science and Leadership courses at the United States Military Academy. The sample was a convenience sample, composed primarily of first- and third-year cadets. Participants were recruited through the department's SONA research participation system and received extra credit for participating. Data were collected anonymously using a Qualtrics-based survey.

After reading each summary individually, participants rated two outcomes: perceived relevance and intention to listen. Perceived relevance was measured using a five-point Likert scale ranging from *Not at all relevant* to *Extremely relevant*. Intention to listen was measured using a parallel five-point scale ranging from *Very unlikely* to *Very likely* to listen to the full episode. Using similar response formats ensured consistent scaling across measures. Because the study employed a within-participant design, each individual evaluated both AI-generated and human-written summaries, allowing participants to serve as their own control. Responses from participants who failed an embedded attention check were excluded prior to analysis.

2.4 Analysis Approach

We used two complementary approaches to evaluate treatment effects. First, we calculated Cliff's delta (δ) to estimate the ordinal effect size between AI-generated and human-generated summaries for each outcome. Because responses were measured on a five-point Likert scale, Cliff's delta provides a distribution-free estimate of how often ratings in the AI condition exceeded ratings in the human condition, minus the reverse comparison. This produces an interpretable measure of effect size and direction. Cliff's delta was used descriptively across all observations, while inferential testing was conducted using mixed-effects models that accounted for clustering at the participant and episode levels.

Second, to address the repeated-measures structure of the data, we estimated linear mixed-effects models (LMMs) using numerically coded Likert responses. Fixed effects included summary condition (AI vs. human), topic domain, Flesch-Kincaid grade level of the summary text, and presentation order. We also included an interaction between summary condition and topic domain to test whether treatment effects differed across content areas. Flesch-Kincaid grade level was included as a continuous covariate to examine whether differences in readability were associated with outcome ratings independent of summary condition. Random intercepts were specified for participants and podcast episodes to account for baseline differences in response tendencies and variation across content.

3. Results

A total of 335 respondents participated in the study. After excluding participants who failed the embedded attention check, the final analytic sample consisted of 304 participants. Approximately 49% were upper-class cadets (Cows and Firsties) and 51% were under-class cadets (Plebes and Yearlings). The sample was predominantly male (72%), with 27% female participants, reflecting typical academy-wide gender distributions rather than a gender-balanced cohort. Participants represented both STEM and non-STEM majors; 55% were enrolled in STEM disciplines, while 45% were enrolled in non-STEM fields, including Behavioral Sciences and Leadership, Social Sciences, Law and Philosophy, History, and English.

ISBN: 97819384962-5-7

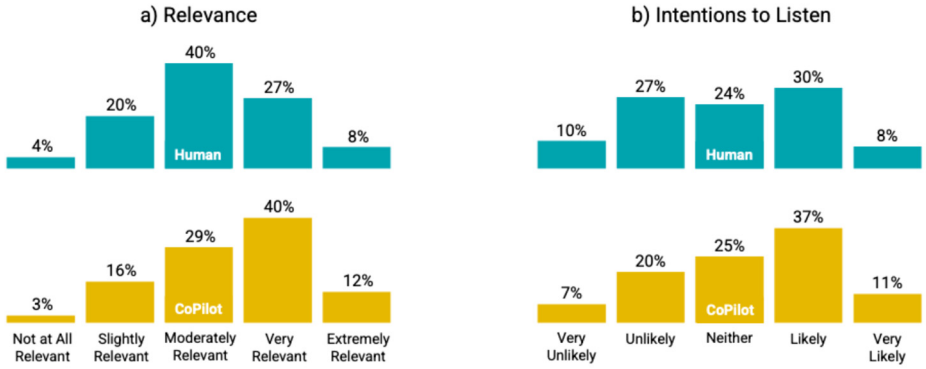

Figure 1. Percentage distribution of Likert-scale responses for (a) perceived relevance and (b) listening intent.

Figure 1 displays the distribution of perceived relevance and intention-to-listen ratings by summarization method. AI-generated summaries (CoPilot) received slightly higher perceived relevance ratings overall relative to human-authored summaries (Cliff's $\delta = .162$, 95% CI [.10, .22]), indicating a small but meaningful ordinal effect. This effect size implies that, across all pairwise comparisons, AI-generated summaries were rated higher than human summaries about 16% more often than the reverse. AI-generated summaries also increased listening intentions (Cliff's $\delta = .125$, 95% CI [.06, .19]), indicating a smaller but directionally consistent effect. As shown in Figure 1a, AI summaries were more concentrated in the *very relevant* category (40%), while human summaries were centered in *moderately relevant* (40%). In Figure 1b, AI-generated summaries yielded a higher proportion of *likely* and *very likely* responses (48% vs. 38%), a +10 percentage point difference. Human summaries were more frequently rated in the lower-intention categories. This alignment between evaluative ratings and reported engagement suggests that differences in perceived relevance translated into differences in downstream listening intentions.

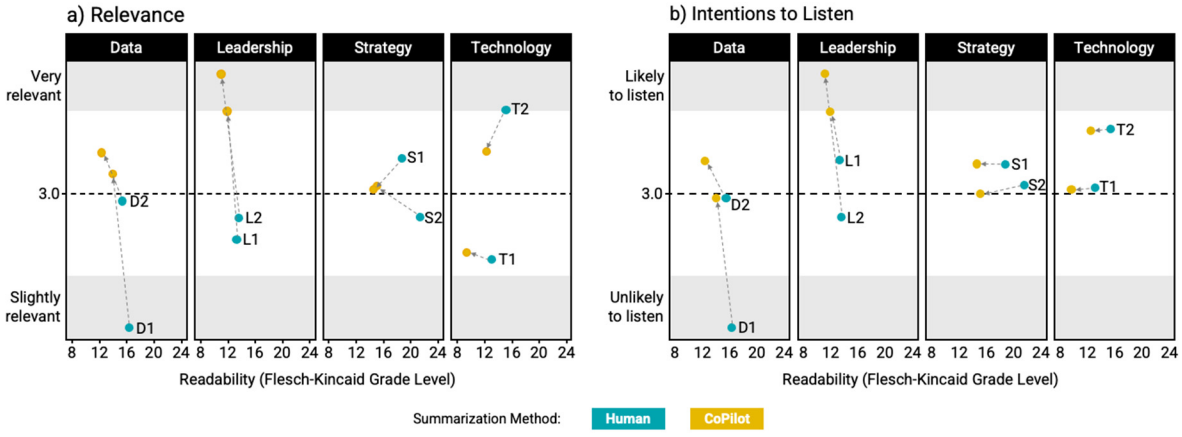

Figure 2. Readability (Flesch–Kincaid grade level) plotted against (a) perceived relevance and (b) listening intent, by topic domain. Points represent individual episode summaries (numeric labels denote episode number). Although AI-generated summaries were generally associated with lower grade-level scores, differences in engagement outcomes varied by domain.

To assess whether these descriptive patterns remained after accounting for repeated measures, topical variation, readability, and order effects, we estimated linear mixed-effects models with random intercepts for participant and episode. Fixed effects included summary method, topic domain, their interaction, Flesch–Kincaid grade level (continuous), and presentation order. For perceived relevance, the model revealed no significant overall main effect of treatment ($\beta = -.31$, SE = .42, $p = .47$), but a significant interaction between treatment and topic. As illustrated in Figure 2a, AI-generated summaries significantly increased relevance ratings for Leadership episodes ($\beta = .86$, SE = .30, $p = .004$) and Data episodes ($\beta = .64$, SE = .24, $p = .008$), relative to the reference topic. No significant treatment effect was observed for Technology episodes. These findings indicate that the effectiveness of AI-generated summaries depends on content domain rather than reflecting a uniform advantage. The model for intentions to listen produced a similar pattern shown in Figure 2b. A significant treatment-by-topic interaction indicated that AI-generated summaries increased listening intentions for Leadership episodes ($\beta = .63$, SE = .32, p

= .05) and Data episodes (β = .55, SE = .26, p = .04), while no significant effect was observed for Technology episodes. Together, these findings demonstrate that AI-generated summarization influences both evaluative judgments and behavioral intentions, but that these effects vary systematically by topical domain.

4. Discussion

The results show that the effects of AI-generated summaries were domain-specific. Relative to the Strategy reference category, AI summaries significantly increased perceived relevance and self-reported listening intentions in Leadership and Data topics. No significant differences were observed in Technology episodes. Importantly, human-written summaries did not outperform AI-generated summaries in any domain. Together, this pattern suggests that automated summarization offers potential benefits without evidence of systematic disadvantage, though one instance of a fabricated direct quote was identified despite constraints to use only source transcripts.

First, we examined readability. AI summaries were easier to read across all episodes. Overall length was similar (AI: M = 113 words; Human: M = 110 words), but AI summaries had lower Flesch–Kincaid Grade Level scores (AI: M = 12.64; Human: M = 15.93) and shorter sentences (AI: M = 14.4 words; Human: M = 24.2 words). This indicates that the AI consistently simplified language and presentation.

Second, we considered whether readability explains the engagement effects. If readability alone explained the effect, gains would be expected across all domains. The absence of improvement in Technology and Strategy episodes suggests that simplifying language is not enough. It is also unlikely that surface cues explain the differences. Podcast titles and speaker information were identical across conditions and were not modified or emphasized differently by the AI. Because these elements were held constant, differences in perceived relevance cannot be attributed to changes in podcast titles or speaker recognition.

These findings set a limit on fluency explanations. Although processing fluency can increase credibility (Alter & Oppenheimer, 2009), clearer writing alone did not guarantee higher engagement. Readability improvements were present across all domains, yet behavioral effects emerged only in Leadership and Data topics. This pattern shows that structural clarity alone is not enough. Domain alignment matters.

We also observed differences in variability across episodes. Human summaries showed a wider range of average ratings, including one notably low-performing episode. AI summaries fell within a narrower range and did not produce comparably low scores. Although AI did not outperform human summaries in every case, it appeared to reduce poor outcomes while improving performance in select domains. In operational settings, avoiding poor outcomes may be as important as raising average performance.

5. Limitations

Although the findings suggest that AI-based summarization can be effective in certain contexts, several limitations shape how broadly the results should be interpreted. The participant sample consisted entirely of United States Military Academy cadets. This population reflects a structured, high-demand decision environment that fits the applied focus of the study. However, it also limits generalizability. Cadets share a common institutional culture, similar academic standards, and aligned career paths. These shared characteristics may influence how they process information and evaluate relevance. Their responses may not reflect how civilian professionals or leaders in decentralized or commercial settings would respond. For this reason, the findings should be viewed as context-specific rather than universally applicable. Replication in more diverse organizational environments would help determine whether the effects extend beyond this setting. The podcast content was professional but not highly technical. AI systems may perform differently when summarizing complex engineering or specialized material. Future research should test whether these findings extend to those settings.

The way AI summaries were generated also reflects a practical implementation rather than an optimized technical benchmark. Summaries were produced using a standardized prompting approach in Microsoft Copilot. Prompts were refined for clarity and consistency, but model parameters were not systematically adjusted, and RAG was not integrated. Prompt design and retrieval systems can affect coherence and factual accuracy. Because these elements were not varied, this study evaluates the real-world performance of a usable AI workflow rather than the maximum capability of the model. Additionally, the human-written summaries were naturalistic rather than experimentally standardized. This may have introduced variation in structure that was not present in the AI condition. Future studies should examine how prompt calibration, retrieval integration, and domain-specific tuning influence perceived relevance, consistency, and downstream engagement. We also did not conduct model-level benchmarking, such as ROUGE or BERTScore, to compare summary quality across conditions. As a result, our findings reflect user responses rather than traditional metrics of semantic alignment or lexical overlap.

6. Conclusion

This study shows that AI-generated podcast summaries can influence perceived relevance and engagement in institutional communication settings. In this context, AI summaries performed at least as well as human-written summaries and showed advantages in certain domains. Automated summaries increased listening intent in multiple episodes and demonstrated more consistent performance across topics, suggesting they can help audiences prioritize information and decide more quickly what is worth their time. At the same time, effectiveness depended on the topic. Structural clarity and readability appeared to play a role, though they were not sufficient to explain the domain-specific pattern. Content selection and alignment with domain-specific expectations also appear to matter. Importantly, AI summaries increased listening intent in some areas without reducing engagement in others, supporting their practical scalability. However, manual inspection identified a fabricated direct quote in one instance, indicating that even constrained prompting does not fully eliminate the risk of factual errors. This suggests that verification remains necessary when outputs are used for communication or decision-making. Future research should test effects across more diverse operational settings and evaluate whether optimized prompt engineering or RAG produces stronger or more consistent engagement outcomes.

7. Acknowledgments

The authors used ChatGPT to assist with editing for clarity, structure, and conciseness. All substantive content, analysis, and interpretations were developed and verified by the authors, who take full responsibility for the final manuscript. This study was approved under USMA HRPP protocol CA-2026-46.

8. References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Belveal, S., Harkins, D., O'Donnell, C., Platt, S., & Koban, D. (2025). Maximizing the usability of publicly available information. *Proceedings of the Annual General Donald R. Keith Memorial Conference*, 7–12. https://www.ieworldconference.org/content/WP2025/Papers/GDRKMCC25_2.pdf

Kirstein, F., Ruas, T., Kratel, R., & Gipp, B. (2024). Tell me what I need to know: Exploring LLM-based (personalized) abstractive multi-source meeting summarization. *arXiv*. https://doi.org/10.48550/arxiv.2410.14545

Klein, L. K., Earl, E., & Cundick, D. (2023, May 1). Reducing information overload in your organization. *Harvard Business Review*. https://hbr.org/2023/05/reducing-information-overload-in-your-organization

Laskar, M. T. R., Fu, X.-Y., Chen, C., & TN, S. B. (2023). Building real-world meeting summarization systems using large language models: A practical perspective. *arXiv*. https://doi.org/10.48550/arXiv.2310.19233

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv*. https://doi.org/10.48550/arxiv.2005.11401

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop on Text Summarization Branches Out* (pp. 74-81).

Liu, Y., Yang, T., Huang, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., & Zhang, Q. (2023). Calibrating LLM-based evaluator. *arXiv*. https://doi.org/10.48550/arxiv.2309.13308

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318).

Peikos, G., & Pasi, G. (2024). A systematic review of multidimensional relevance estimation in information retrieval. *WIREs Data Mining and Knowledge Discovery*, *14*(5), e1541. https://doi.org/10.1002/widm.1541

Samimi, P., & Ravana, S. D. (2014). Creation of reliable relevance judgments in information retrieval systems evaluation experimentation through crowdsourcing: A review. *The Scientific World Journal*, *2014*, 1–13. https://doi.org/10.1155/2014/135641

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *arXiv*. https://doi.org/10.48550/arxiv.1904.09675