



Agentic Artificial Intelligence for the Automated Generation of Accurate Summary Podcasts of Radiology Research Papers

Jeong Hyun Lee¹, Kyowon Gu¹, Jaeseung Shin¹, Ji Hye Min¹, Jeong Ah Hwang¹, Ji Yoon Ban¹, Jihwan Choi¹, Taewon Han¹, Dong Young Jeong¹, Kyung A Kang¹, Haejung Kim¹, Harim Kim¹, Hyun Su Kim^{1,2}, Min Je Kim^{1,3}, Myoung Kyoung Kim¹, Sanghee Kim¹, Min Hwan Kwak⁴, Ji Hyun Lee¹, Sang Yub Lee¹, Chang Hoon Oh¹, Sung Goo Park¹, Yun Hwa Roh^{1,5}, Yongsik Sim¹, Beomseok Sohn¹, Siyoun Sung⁶, Min-Hyuk Yu¹, Seong Ho Park⁵

¹Department of Radiology and Center for Imaging Sciences, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

²Musculoskeletal Imaging and Intervention, Department of Radiology, University of Washington, Seattle, WA, USA

³Department of Radiology, U-Well Urology Clinic, Seoul, Republic of Korea

⁴Department of Radiology, Armed Forces Goyang Hospital, Goyang, Republic of Korea

⁵Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea

⁶Department of Radiology, Gidaechan Hospital, Gwangmyeong, Republic of Korea

Objective: To evaluate whether a custom agentic artificial intelligence (AI) pipeline can overcome the limitations of general-purpose large language model tools, when compared with a generic commercial tool (Google NotebookLM [NBLM]), for generating podcast-style summaries of radiology research articles.

Materials and Methods: Twenty-two PDF-format original research articles published in the April 2025 issue of *Radiology* were processed using our Programmable, Phoneme-Aware PDF-to-Podcast Pipeline (P5) and NBLM to generate 44 audio episodes. P5 utilizes a multi-agent workflow for script generation, quality assurance, pronunciation enhancement, and audio synthesis. Four radiologists from a pool of 25 (7 generalists and 18 specialists) were randomly assigned to evaluate each blinded audio episode, yielding 176 total evaluations. The primary outcomes were the number of hallucinations (factual errors) per episode and the percentage of hallucination-free episodes. Secondary outcomes included the number of inappropriate statements, mispronunciations, and flow disruptions; the composite quality score (Quality Assessment of Educational Podcasts [QAEP]); the key results coverage score; and overall listener preference. Data were analyzed using generalized linear mixed models.

Results: The P5 method produced significantly fewer hallucinations per episode compared with NBLM (mean, 0.32 vs. 0.93; $P < 0.001$) and a higher proportion of hallucination-free episodes (71.6% [63/88] vs. 56.8% [50/88]; $P = 0.013$), consistently across generalists and specialists. P5 demonstrated significantly fewer mispronunciations (mean, 0.11 vs. 1.62; $P < 0.001$) and flow disruptions (mean, 0.19 vs. 1.06; $P < 0.001$) per episode, as well as higher mean QAEP composite scores (4.62 vs. 4.28; $P < 0.001$), compared with NBLM, consistently across generalists and specialists. Raters preferred P5 over NBLM in 72.7% (64/88) of comparisons ($P = 0.003$).

Conclusion: Our custom agentic AI pipeline generated podcast-style summaries of radiology research articles with significantly higher quality and greater listener preference than the generic commercial tool.

Keywords: Artificial intelligence; Large language model; Generative; Podcast; Audio summary; Agent; Radiology; Automation

Received: December 10, 2025 **Revised:** March 14, 2026 **Accepted:** March 19, 2026

Corresponding author: Seong Ho Park, MD, PhD, Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea

• E-mail: parksh.radiology@gmail.com

Current affiliations: Hyun Su Kim—Musculoskeletal Imaging and Intervention, Department of Radiology, University of Washington, Seattle, WA, USA; Min Je Kim—Department of Radiology, U-Well Urology Clinic, Seoul, Republic of Korea; Yun Hwa Roh—Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea; Min-Hyuk Yu—Department of Teleradiology, Armed Forces Capital Hospital, Seongnam, Republic of Korea

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

See the invited Editorial “Generative Artificial Intelligence for Empowering Non-Native English Speakers in Verbal Scientific Communication and Dissemination” at <https://doi.org/10.3348/kjr.2026.0474>.

INTRODUCTION

Podcasts have become an important vehicle for delivering professional medical information, offering busy clinicians an efficient and flexible means of engaging with current medical knowledge [1,2]. The audio-only format facilitates learning during commuting or exercise, making podcasts particularly attractive for continuing medical education. High-quality medical podcasts can deliver knowledge comparable to that of conventional reading formats, and studies have shown that many physician listeners modified their clinical practice based on podcast content [2].

Journal-affiliated podcasts, including the *Radiology* podcast series, play an important role in disseminating published research. Recent evidence has demonstrated that featuring articles on journal podcasts enhances their academic impact, with measurable increases in downloads, citations, and Altmetric Attention Scores [1]. However, producing high-quality podcasts remains time- and labor-intensive.

With advances in large language models (LLMs), researchers have explored the use of artificial intelligence (AI) to automate podcast creation from academic articles [3]. Several journals—including the *Bone & Joint Journal*, the *British Journal of Dermatology*, the *Journal of the American Medical Association*, the *Korean Journal of Radiology*, and *Radiology: Artificial Intelligence*—have begun experimenting with LLMs for this purpose by directly offering AI-generated audio summaries or by using LLMs to create content summaries and multimedia transcripts [4-8].

Among the tools available for this purpose, Google NotebookLM (NBLM)—an AI-powered research assistant that generates podcast-style audio summaries termed “Audio Overviews”—has gained particular attention since its public launch in 2024, garnering millions of monthly visits [9]. However, recent research suggests that such general-purpose tools may have significant limitations when applied to complex clinical research articles, such as those in the medical imaging field [3]. Tejani et al. [3] evaluated NBLM’s Audio Overview feature using 21 articles published in *Radiology* and found that 71% of the podcasts contained

factual errors (i.e., hallucinations), and that the content was overly simplified for specialists. These findings suggest that general-purpose AI tools may not meet the accuracy standards required by medical professionals without further refinement.

To address this limitation, we developed a custom pipeline designed to generate more accurate AI-produced podcast-style audio summaries from radiology research papers. This study evaluated whether our custom pipeline could produce higher-quality podcast-style AI-generated audio episodes than NBLM and whether these quality improvements would be consistently recognized by radiologists with different levels of expertise.

MATERIALS AND METHODS

Ethics and Copyright Protection

This nonhuman-subject experimental study did not require institutional review board approval. The study used AI tools under an AI training opt-out policy, ensuring that no proprietary or copyrighted source materials were retained or used for model development. To maintain strict copyright and privacy standards, AI-generated audio files were distributed exclusively to designated evaluators through secure access points. All generated assets were permanently deleted upon completion of the study to prevent potential data leakage or unauthorized use. Permission to use *Radiology* articles in the AI tools in this study was granted by the Radiological Society of North America.

Study Design

This study employed a within-subject experimental design to compare our custom Programmable, Phoneme-Aware PDF-to-Podcast Pipeline (P5) with NBLM for generating podcast-style summaries of radiology research articles written in English. We selected NBLM as the benchmark because it was previously evaluated specifically for generating podcast-style audio summaries of radiology research articles [3], in addition to its prominence among similar software tools. The study code is available at <https://github.com/sigjhl/p5>. This study adheres to the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM) checklist (Supplementary Table 1) [10].

Agentic Podcast Generation Pipeline

We developed an agentic pipeline to convert research papers from PDF format into podcast-style summary audio

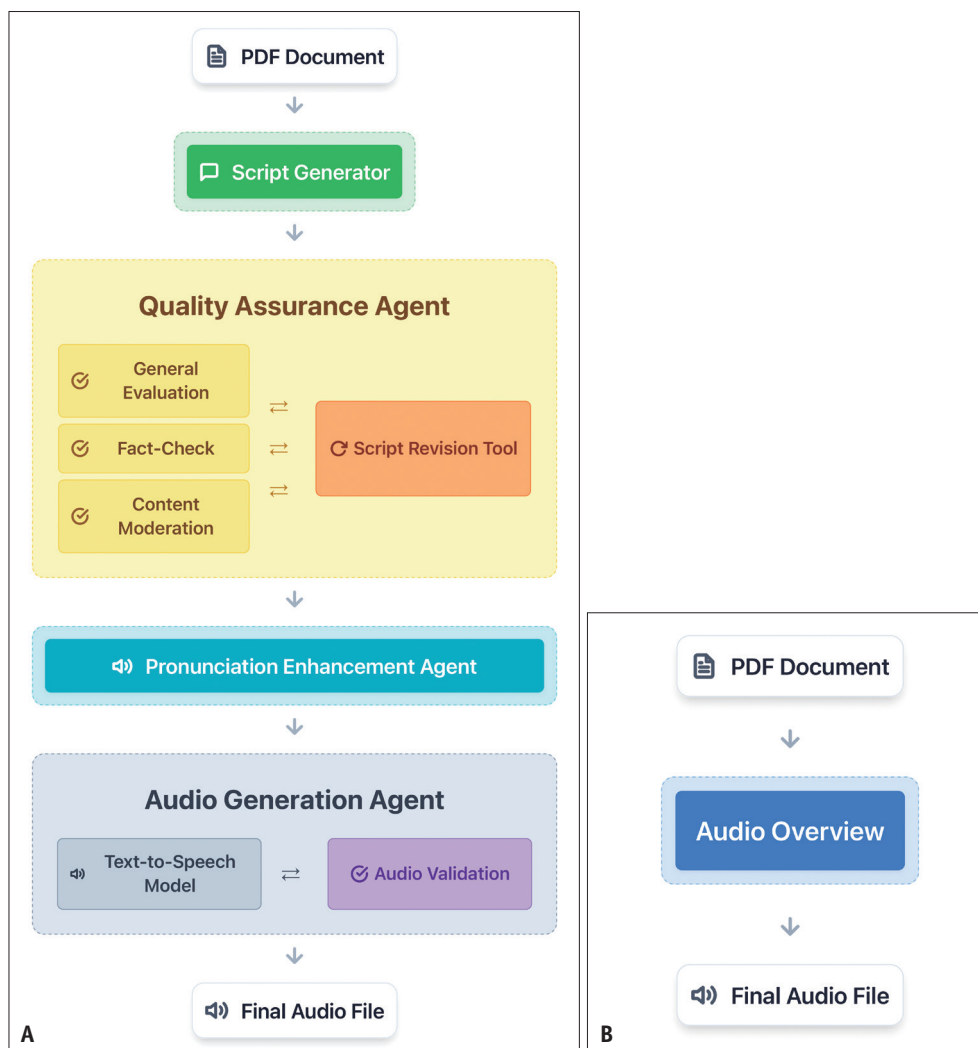


Fig. 1. Schematic diagrams depicting the process of audio generation. **A, B:** Panel **(A)** illustrates audio generation using our custom method (P5), and Panel **(B)** illustrates audio generation using the generic method (NBLM). P5 = Programmable, Phoneme-aware PDF-to-Podcast Pipeline, NBLM = NotebookLM

episodes using AI (Fig. 1). Our pipeline, P5, employs AI agents that perform specific functions at multiple stages, utilizing gemini-2.5-pro (Google) for script generation and quality assurance, o3-2025-04-16 (OpenAI) for pronunciation enhancement, and gemini-2.5-pro-preview-tts (Google) for audio synthesis. All models were accessed via application programming interfaces. Further technical details are provided in Supplement (Technical Implementation Details of the P5 Pipeline)

The pipeline incorporates three key components designed to address common limitations of general-purpose AI tools. A quality assurance agent evaluates generated scripts through multi-step verification, including extraction and verification of factual claims against the source document, assessment of content appropriateness, and revision

of scripts based on identified issues. A pronunciation enhancement agent analyzes scripts to identify medical terms that are likely to be mispronounced by text-to-speech systems and then inserts appropriate phonetic annotations. An audio generation agent converts phoneme-enhanced scripts into audio, performs quality assurance of the outputs, and automatically regenerates audio when issues are detected. Each PDF was uploaded to gemini-2.5-pro for persistent access, allowing the agents to reference the document's text and visual elements throughout the generation workflow. Details regarding PDF document processing and content access during generation, the configurations used for script generation, and the roles of the three agents are provided in Supplement (Technical Implementation Details of the P5 Pipeline).

Study Materials

All original research articles published in the April 2025 issue of *Radiology* served as source materials; the complete list of articles is provided in Supplementary Table 2. Between July 12 and 18, 2025, a pair of AI-generated audio episodes was generated for each paper, one using P5 and the other using NBLM. Both systems received identical custom instructions emphasizing accurate coverage of key results, appropriate technical depth for radiologist audiences, and an engaging conversational style (Fig. 2). The same PDF files were uploaded to both systems without modification. All generations for both methods were completed successfully without manual intervention.

Study Participants as Raters

The evaluation was conducted in a randomized, blinded manner. Twenty-five radiologists, who were fluent in English, participated as raters, comprising 7 generalists and 18 specialists. Generalists included radiology residents and board-certified radiologists without fellowship or subspecialty training. Specialists included radiology fellows and attending radiologists with subspecialty expertise. We included a relatively large number of raters (25) rather than adopting the typical design in radiology research involving a small number of raters with similar characteristics. This approach was intended to further mitigate the inherently subjective nature of the evaluation outcomes, beyond the incorporation of a standardized pre-study rater training procedure (explained later), thereby enhancing the reliability of the evaluations. To minimize bias arising from prior experience or familiarity with audio recordings generated by either P5 or NBLM, we limited participation to individuals with less than one hour of experience with

NBLM-generated audio. None of the participants had prior experience with audio recordings generated by P5.

Study Procedures

Each radiologist was assigned to evaluate AI-generated podcast-style summary audio episodes for a subset of papers (one to seven papers per radiologist), with random assignment except for subspecialty matching among specialists. Each paper and its associated summary audio episodes were assessed by four radiologists (two generalists and two specialists matched to the paper's subspecialty), randomly selected from the 25 raters. This prespecified random rater allocation was chosen to maintain a manageable per-rater listening burden while allowing each episode to be evaluated from multiple independent perspectives drawn from a broader and more heterogeneous rater pool, thereby enabling a more balanced comparison between P5 and NBLM. For each assigned paper, raters were provided with two audio recordings in a single evaluation session—one generated by P5 and the other by NBLM. Raters were blinded to the specific AI tool used for each recording, and the presentation order of the P5 and NBLM versions was randomized for each paper and for each rater to minimize evaluation bias. To reflect typical real-world use, raters were permitted to consult the PDF before, during, or after listening and were allowed to pause or rewind the audio as needed.

We adopted this evaluation scheme rather than one involving a washout period to control for carryover effects, such as in a crossover design [11], which is often used for comparisons of paired image sets. This decision was based on the fact that one outcome of this study directly assessed preference between the two recordings, thereby necessitating evaluation of both within the same session (as explained later). In addition, potential carryover effects were considered less influential in our study because the constant element was the underlying PDF article rather than the two generated audio summaries. This differs from paired image comparisons, in which the two images are essentially identical except for the specific technical modifications under evaluation. The most critical measure for bias control in our study was rater blinding to the AI tools, which was rigorously implemented.

Raters evaluated episodes using standardized rating criteria for which they had received prior remote training. The training used materials not included in the study evaluation and consisted of distribution of a written rating guide

Your goal is to transform a dense research paper into an engaging, insightful, and educational podcast script aimed at practicing radiologists and radiologists-in-training.

If the article includes a "Key Results" section, ensure every item in that section is conveyed faithfully.

Deepen the script with probing questions about applications, limitations, and future directions. Present radiological concepts accurately and with the technical depth suitable for professionals, allowing clarity and curiosity—rather than clinical authority—to carry the conversation.

Fig. 2. Instructions for generating the initial script for P5 and the audio summary for NBLM. P5 = Programmable, Phoneme-aware PDF-to-Podcast Pipeline, NBLM = NotebookLM

developed by the lead author (J.H.L.), which defined each evaluation item with concrete examples (e.g., hallucination: reporting an area under the curve of “0.93” vs. “0.83” in the source paper), followed by a single round of rater questions and feedback with corresponding revision of the guide, and a recorded walkthrough reinforcing the evaluation criteria.

For each audio episode, raters first provided free-text descriptions of identified issues and then submitted category counts (hallucinations, inappropriate statements, mispronunciations, and flow disruptions). A hallucination was recorded when an episode included a factual statement that contradicted the source PDF or asserted a specific fact not supported by it, including fabricated content or numerical inaccuracies. Inappropriate statements were defined as remarks unsuitable for a professional medical education context, and flow disruptions were defined as interruptions that materially degraded the listening experience. Raters also scored key results coverage (0 = none, 0.5 = partial, 1.0 = full). Raters then completed structured quality ratings using the Quality Assessment of Educational Podcasts (QAEP) instrument [2,12] with a radiology-adapted, modified QAEP scoring form (5-point scale; higher scores indicate better quality) [3]. The modified 20-item QAEP instrument assessed the quality of the generated audio episodes across multiple dimensions, including audio clarity, conversational naturalness, content accuracy, educational value, and appropriateness for radiologist audiences. Finally, raters indicated their overall preference between the two audio recordings for each paired comparison. Further details, including detailed definitions, counting rules, examples, and the full rating guide, are provided in Supplement (Rater Training and Evaluation Protocol).

Outcome Measures

The primary outcomes were the mean number of hallucination events per audio episode and the percentage of audio episodes that were free of hallucinations.

Secondary outcomes included: 1) the number of inappropriate statements per audio episode (any remarks violating professional or sensitivity standards), 2) the number of mispronunciations per audio episode (medical or radiological terms only; repeated or multiple errors on the same word were counted once), 3) the number of flow disruptions per audio episode (irrelevant dialogue, third-party voice intrusion, or significant background noise), 4) the composite Likert score (mean of the 20 modified QAEP items), 5) the key results coverage score (mean

completeness across key results items: 0 = no coverage, 0.5 = partial coverage, 1.0 = full coverage; papers without key results were excluded), and 6) overall preference between the two audio episodes. Additionally, individual QAEP items were analyzed separately to identify the specific quality dimensions driving the overall differences.

Statistical Analysis

All analyses were conducted using generalized linear mixed models (GLMMs) to account for the paired design and repeated evaluations by four readers per AI-generated audio episode. We fitted negative binomial (log link), Gaussian (identity link), and binomial (logit link) GLMMs for count, continuous, and binary outcomes, respectively. All models included fixed effects for AI method (P5 vs. NBLM), rater type (generalist vs. specialist), and listening order, with random effects for paper, rater, and the paper x rater interaction. An additional analysis tested the interaction between AI method and rater type using likelihood ratio tests. False discovery rate (FDR) correction using the Benjamini–Hochberg procedure was applied to account for increased type I error due to multiple outcome comparisons, separately for: 1) primary outcomes (two tests), 2) secondary outcomes (six tests), and 3) individual QAEP items (20 tests), with statistical significance defined as FDR-adjusted $P < 0.05$. Additionally, generation time and audio duration were compared between methods using a paired t -test. All analyses were conducted in R version 4.5.1 using lme4 (version 1.1-37), MASS (version 7.3-61), emmeans (version 1.11.2-8), and tidyverse (version 2.0.0), with maximum likelihood estimation.

RESULTS

A total of 22 articles were used as source materials for podcast generation. This process yielded 44 generated audio episodes. Because four evaluations were performed for each article (two by generalists and two by specialists), the total number of evaluations was 176, comprising 88 paired preference judgments between P5 and NBLM (Table 1).

Primary Outcomes

Rater type did not influence any primary outcomes ($P > 0.05$), and the results were consistent between generalist and specialist raters. Podcasts generated using the P5 method demonstrated a significantly lower mean number of hallucination events per episode than those generated

using NBLM (0.32 vs. 0.93; $P < 0.001$), representing a 66% reduction (Table 2). The P5 method produced a significantly higher proportion of podcast audio episodes that were entirely free of hallucinations (71.6% [63/88] vs. 56.8% [50/88]; $P = 0.013$) (Table 2).

Table 1. Study sample, rater, and subspecialty distribution

Item	Value
Raters*	
Generalist	7 (28.0)
Specialist	18 (72.0)
Articles evaluated per rater [†]	3 [1–7]
Materials for podcast generation and evaluation	
Number of articles used	22
Number of generated podcasts	44
Number of total evaluations	176
Paired preference evaluations	88
Raters per article	4 (2 generalists, 2 specialists)
Source article subspecialty [‡]	
Breast imaging	4 (18.2)
Cardiothoracic	1 (4.5)
Computer applications and physics	3 (13.6)
Gastrointestinal	3 (13.6)
Genitourinary	1 (4.5)
Interventional radiology	1 (4.5)
Musculoskeletal	3 (13.6)
Neuroradiology	6 (27.3)

Values are n (%) unless otherwise indicated.

*Percentages out of all raters (n = 25), [†]Presented as median

[range], [‡]Percentages out of all papers (n = 22)

Secondary Outcomes

Rater type did not influence any secondary outcomes ($P > 0.05$) except for the number of inappropriate statements. A significant interaction was observed between rater type and AI method for the number of inappropriate statements ($P = 0.034$). Among generalist raters, the mean number of inappropriate statements was significantly lower for P5 (0.14 vs. 0.50; $P = 0.005$), whereas no significant difference was observed between P5 and NBLM among specialist raters (0.11 vs. 0.07; $P = 0.483$) (Table 2).

For the remaining five secondary outcomes, results were consistent across generalists and specialists. The P5 method demonstrated significantly superior results compared with NBLM for all measures except key results coverage (Table 2). The P5 method showed a significantly lower mean number of mispronunciations (0.11 vs. 1.62; $P < 0.001$) and flow disruptions (0.19 vs. 1.06; $P < 0.001$) per audio episode than NBLM. The subjective quality assessment also favored P5. The mean QAEP composite score was significantly higher for P5 (4.62 vs. 4.28; $P < 0.001$). Raters demonstrated a strong overall preference for the P5 podcasts (72.7% [64/88] vs. 27.3% [24/88]; $P = 0.003$). Additional exploratory comparisons between P5 and NBLM across the 20 individual QAEP items revealed significantly higher scores for P5 in 13 of the 20 items (all $P < 0.05$) (Fig. 3).

Generation Time

The P5 method required significantly more computational time than NBLM (Table 3). The mean generation time for

Table 2. Comparison of P5- and NBLM-generated audio episodes

Outcomes	P5	NBLM	FDR-adjusted P
Primary outcomes			
Hallucinations per audio episode	0.32 ± 0.56	0.93 ± 1.42	<0.001
Hallucination-free audio episodes	71.6 (63/88)	56.8 (50/88)	0.013
Secondary outcomes			
Inappropriate statements per audio episode*			
Generalist raters	0.14 ± 0.55	0.50 ± 1.30	0.005
Specialist raters	0.11 ± 0.39	0.07 ± 0.33	0.483
Mispronunciations per audio episode	0.11 ± 0.38	1.62 ± 1.90	<0.001
Flow disruptions per audio episode	0.19 ± 0.48	1.06 ± 1.80	<0.001
QAEP composite score	4.62 ± 0.47	4.28 ± 0.70	<0.001
Key results coverage	0.98 ± 0.06	0.96 ± 0.13	0.154
Overall preference: P5 vs. NBLM	72.7 (64/88)	27.3 (24/88)	0.003

Values are mean ± standard deviation or percentage (numerator/denominator).

*Results are provided separately for generalist and specialist raters due to a significant interaction between rater type and artificial intelligence methods ($P = 0.034$).

P5 = Programmable, Phoneme-aware PDF-to-Podcast Pipeline, NBLM = NotebookLM, FDR = false discovery rate, QAEP = Quality Assessment of Educational Podcasts

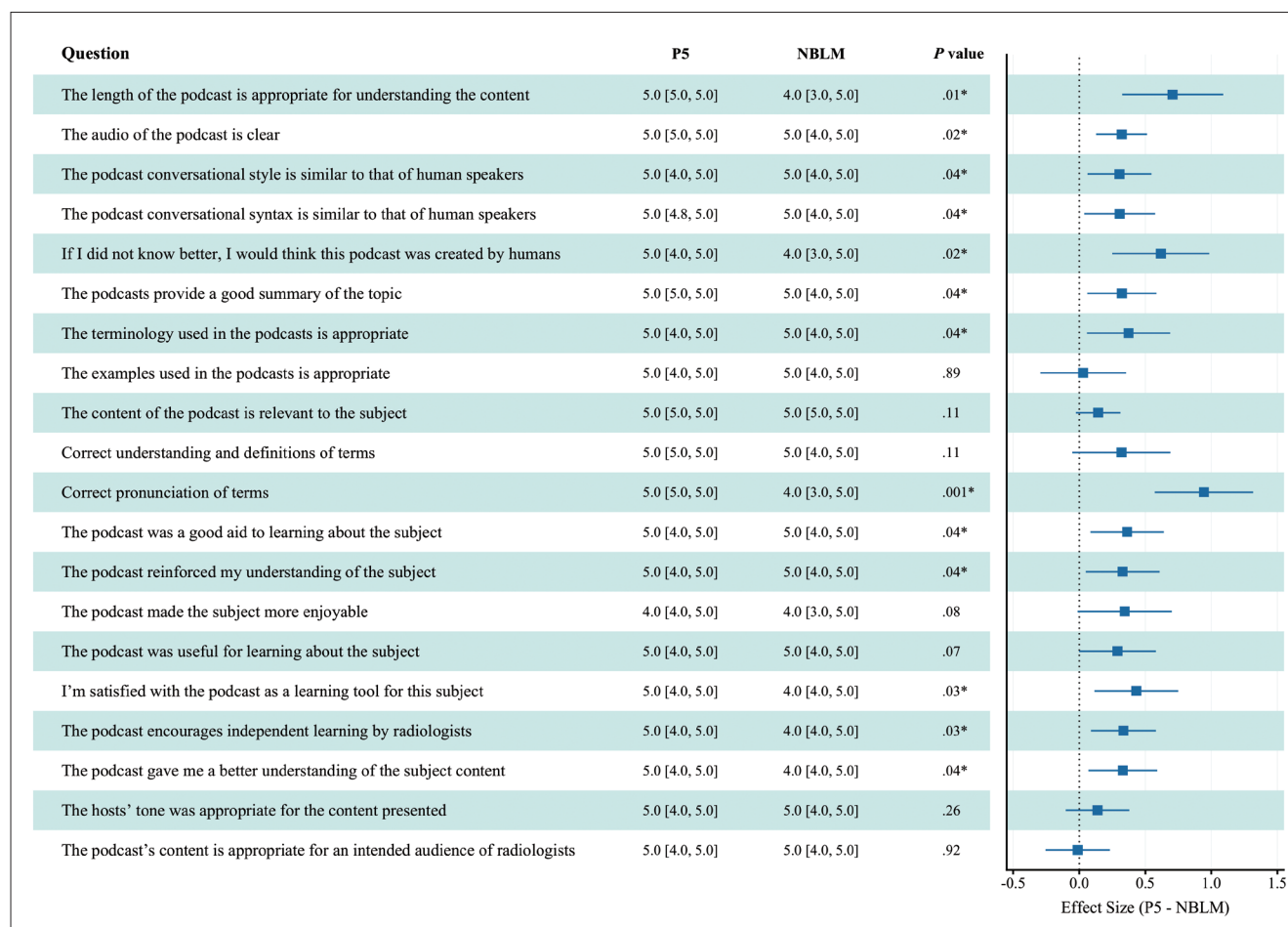


Fig. 3. Comparison of individual items included in the QAEP Likert scores between P5 and NBLM. The forest plot shows the estimated effect sizes for all 20 QAEP questionnaire items. The central dot represents the point estimate of the effect size (mean difference), and the horizontal line denotes its 95% confidence interval. The vertical reference line at zero indicates no difference between the two models. Positive effect sizes indicate higher scores for P5. The P5 and NBLM columns display median [Q1, Q3] for each method. *P*-values are adjusted for the false discovery rate using the Benjamini–Hochberg procedure. QAEP = Quality Assessment of Educational Podcasts, P5 = Programmable, Phoneme-aware PDF-to-Podcast Pipeline, NBLM = NotebookLM

Table 3. Comparison of resource requirements and audio output length

Item	P5	NBLM	<i>P</i>
Generation time, min	11.4 ± 3.5	3.3 ± 0.3	<0.001
Audio length, min	16.2 ± 1.5	17.7 ± 3.1	0.062

Values are shown as mean ± standard deviation.

P5 = Programmable, Phoneme-aware PDF-to-Podcast Pipeline, NBLM = NotebookLM

P5 was 11.4 minutes, compared with 3.3 minutes for NBLM ($P < 0.001$), despite producing audio episodes of similar duration (16.2 minutes for P5-generated podcasts vs. 17.7 minutes for NBLM-generated podcasts; $P = 0.062$).

DISCUSSION

This study demonstrated that a custom, agentic AI pipeline can produce podcast-style audio summaries of radiology research papers with significantly higher quality than those produced by a generic AI tool. Compared with NBLM, P5 significantly reduced the number of hallucinated statements by two-thirds (mean, 0.32 vs. 0.93 per episode), increased the proportion of error-free podcasts (71.6% vs. 56.8%), and achieved higher scores on the modified QAEP composite (mean, 4.62 vs. 4.28). Mispronunciations and flow disruptions were also significantly fewer, and listeners preferred the episodes generated by P5 in nearly three-quarters of the cases. These benefits were consistent across generalists and specialists.

One goal of scientific journal podcasts is to enhance author engagement—often through interviews that capture behind-the-scenes insights and personal reflections—and to promote deeper academic discourse through expert commentary or panel discussions. Current AI systems cannot replicate this depth of engagement. Consequently, the podcast-style audio summaries generated in this study fall short of the depth and character of a genuine, human-driven journal podcast. Nevertheless, our findings suggest that this approach is feasible as an automated method for generating more accurate summaries that may serve as an effective tool for disseminating new scientific knowledge from research publications.

The superior performance of P5 is likely attributable to its multi-agent architecture. A dedicated quality assurance agent cross-references every scripted claim with the original PDF, while a pronunciation enhancement agent supplies phonetic cues that prevent the text-to-speech engine from mispronouncing radiologic terminology. The audio generation agent, in addition to generating the audio file, screens for issues and determines whether the audio should be regenerated. By embedding these checks directly into the generation workflow, P5 addresses shortcomings previously reported for generic AI tools used for podcast-style audio generation. These additional quality control steps extended the generation time compared with NBLM (mean, 11.4 vs. 3.3 minutes). However, this technical cost does not diminish the advantages of the P5 approach.

Unlike AI applications in time-sensitive use cases, such as detecting intracranial hemorrhage on head CT, academic journal podcast production is not time-critical. Instead, the accuracy and adequacy of the audio content are of primary importance.

Tejani et al. [3] previously evaluated the default “Audio Overview” feature of NBLM for podcast-style audio generation using 21 articles published in *Radiology*. Our NBLM results differed numerically from theirs, particularly in the proportion of hallucination-free audio episodes (56.8% vs. 28.6%) and in audio duration (mean, 17.7 minutes vs. 9.1 minutes). These discrepancies may be attributed to 1) the custom prompt we supplied, which emphasizes accurate coverage of key results and an audience-appropriate level of detail, 2) the different set of articles used (April 2025 vs. September 2024 issues of *Radiology*), and 3) interim updates to the NBLM model between the two studies.

While our pipeline substantially reduced quality issues, it did not eliminate them entirely—reflecting an inherent challenge in generative AI (Table 4). Minor discrepancies in fine details or nuanced interpretations may still occur occasionally. This finding underscores the importance of a human-in-the-loop approach in AI-driven journal podcast production, analogous to the necessity of human supervision in other areas of AI assistance in scientific publication [13,14]. Furthermore, as suggested by the discrepant counts of inappropriate statements per audio episode

Table 4. Examples of identified issues

Type	Issue	Source article	Method
Hallucinations	Key results #2 misquoted: the paper reports 0.5 ± 0.03 vs. 0.48 ± 0.05 , but the podcast says 0.5 ± 0.03 vs. 0.5 ± 0.05	Bøgh et al.*	NBLM
	Incorrectly states the count of annotated cases in ISLES 2022 as 400: the original text states 250 annotated DWI cases and 150 without annotations (for validation/test), not 400 annotated	Jiang et al.†	NBLM
	The volumes were reversed: the podcast says True 3T = 482.1 cm^3 and LowGAN = 483.6 cm^3 , but the paper reports 3T = 483.6 and LowGAN = 482.1	Lucas et al.‡	P5
Inappropriateness	The host says, “It seems like an interesting place for an injury to show up.” In the context of a study on soldiers with brain injuries, this phrasing can sound inappropriately casual	Diociasi et al.§	P5
Mispronunciation	Mispronounced “Gl _x ” as D-L-X	Bøgh et al.*	NBLM
	Mispronounced “nnU-NET” as N-U-NET, “mTICI” as M-T-I-C-I	Jiang et al.†	NBLM
Flow disruption	At 01:58, as the host says “Okay, so if we do want to visualize metabolism specifically,” a third-party voice (‘huh’) overlaps the audio	Liu et al.¶	NBLM

*<https://doi.org/10.1148/radiol.241808>, †<https://doi.org/10.1148/radiol.240775>, ‡<https://doi.org/10.1148/radiol.233529>, §<https://doi.org/10.1148/radiol.233264>, ¶<https://doi.org/10.1148/radiol.241597>.

NBLM = NotebookLM, P5 = Programmable, Phoneme-aware PDF-to-Podcast Pipeline

between generalist and specialist raters, subjective aspects of audio summary quality may require further tailoring to meet the expectations of the primary target audience, for which human oversight may play a more critical role. In this context, the stepwise architecture of the P5 system, which facilitates human modification, represents a distinct advantage over general-purpose tools.

Our study was not intended to establish the practical utility of the P5 system. Ultimately, such utility would be reflected in enhanced knowledge dissemination and its subsequent positive impact on radiology practice and research. However, these theoretically desirable outcomes are difficult to measure directly. Unlike studies assessing more immediate educational effects—such as the impact of different teaching approaches on students' test performance—the influence of research publications on professional practice patterns and knowledge levels is indirect and subject to numerous confounding factors. Consequently, the practical utility of journal podcast-based dissemination is often inferred from surrogate indicators, such as article access, downloads, or citation metrics [1]. Collecting data for such assessments requires long-term collaboration with academic journals, which was beyond the scope of the present study but represents an important direction for future research.

This study had several limitations. Although statistically significant differences between P5 and NBLM were detected across multiple outcomes, the dataset—empirically selected without a formal sample size calculation—remained small for individual subspecialty areas and lacked diversity. Therefore, the results should be interpreted with caution. In addition, all included articles were drawn from a single journal (*Radiology*). We selected *Radiology* because it is the top-ranked journal in the field of radiology and medical imaging, is known for publishing highly sophisticated research, and enables a more relevant comparison with an existing study [3]. Furthermore, our evaluation was limited to English-language content. Although English is the dominant language of scientific publishing [15,16], generalizability to other languages remains uncertain. We did not perform a formal ablation study of the P5 pipeline components or compare P5 with a minimal single-agent Gemini 2.5 Pro baseline to isolate the contribution of the multi-agent architecture. Isolating the multi-agent contribution would be challenging because pronunciation enhancement required the o3 model, and the text-to-speech system itself contributed to audio quality issues

independent of script quality. However, NBLM utilizes Gemini 2.5 Pro for Audio Overview generation according to Google's documentation [17]. Whether it operates as a minimal single-agent wrapper or as its own agentic architecture, our comparison effectively benchmarks P5 against a Gemini 2.5 Pro-based system, with the observed differences attributable to P5's design choices. Additionally, although raters were limited to those with less than one hour of NBLM experience, prior exposure to its characteristic voice may have introduced subtle recognition bias. Finally, as discussed above, direct clinical or educational utility was not assessed.

In conclusion, a custom agentic AI pipeline can generate podcast-style audio summaries of radiology research papers with fewer factual errors and higher overall quality than a generic AI tool. These improvements, driven by specialized agents for quality assurance, pronunciation enhancement, and audio generation, were consistently recognized by listeners across different levels of expertise. This agent-based approach may provide a potential pathway toward more reliable, automated, and scalable production of audio summaries of scientific articles for medical professionals.

Supplement

The Supplement is available with this article at <https://doi.org/10.3348/kjr.2026.0203>.

Availability of Data and Material

The study code is available at <https://github.com/sigjhl/p5>.

Conflicts of Interest

Jeong Hyun Lee: Assistant to the Editor of the *Korean Journal of Radiology*, was not involved in the editorial evaluation or decision to publish this article.

Seong Ho Park: Editor-in-Chief without involvement in the editorial evaluation or decision to publish this article; honoraria from Bayer and Korean Society of Radiology; honoraria and support for travel from Korean Society of Radiology; consulting fees from Ministry of Food and Drug Safety of the Republic of Korea; consulting fees from National Evidence-based healthcare Collaborating Agency of the Republic of Korea.

The remaining authors have declared no conflicts of interest.

Author Contributions

Conceptualization: Jeong Hyun Lee, Seong Ho Park. Data

curation: Jeong Hyun Lee. Formal analysis: Jeong Hyun Lee. Funding acquisition: Jeong Hyun Lee, Jaeseung Shin. Investigation: Kyowon Gu, Jaeseung Shin, Ji Hye Min, Jeong Ah Hwang, Ji Yoon Ban, Jihwan Choi, Taewon Han, Dong Young Jeong, Kyung A Kang, Haejung Kim, Harim Kim, Hyun Su Kim, Min Je Kim, Myoung Kyoung Kim, Sanghee Kim, Min Hwan Kwak, Ji Hyun Lee, Sang Yub Lee, Chang Hoon Oh, Sung Goo Park, Yun Hwa Roh, Yongsik Sim, Beomseok Sohn, Siyoun Sung, Min-Hyuk Yu. Methodology: Jeong Hyun Lee. Project administration: Jeong Hyun Lee, Seong Ho Park. Software: Jeong Hyun Lee. Supervision: Seong Ho Park. Visualization: Jeong Hyun Lee. Writing—original draft: Jeong Hyun Lee, Kyowon Gu. Writing—review & editing: Seong Ho Park, Jaeseung Shin.

ORCID IDs

Jeong Hyun Lee
<https://orcid.org/0000-0002-7125-8899>
Kyowon Gu
<https://orcid.org/0000-0003-2494-018X>
Jaeseung Shin
<https://orcid.org/0000-0002-6755-4732>
Ji Hye Min
<https://orcid.org/0000-0002-8496-6771>
Jeong Ah Hwang
<https://orcid.org/0000-0002-8012-995X>
Ji Yoon Ban
<https://orcid.org/0009-0008-6545-1044>
Jihwan Choi
<https://orcid.org/0009-0009-5214-0244>
Taewon Han
<https://orcid.org/0000-0003-0243-1608>
Dong Young Jeong
<https://orcid.org/0000-0003-4338-131X>
Kyung A Kang
<https://orcid.org/0000-0003-2502-1206>
Haejung Kim
<https://orcid.org/0000-0003-4855-9711>
Harim Kim
<https://orcid.org/0000-0002-0664-4216>
Hyun Su Kim
<https://orcid.org/0000-0002-0179-9542>
Min Je Kim
<https://orcid.org/0000-0003-3672-7052>
Myoung Kyoung Kim
<https://orcid.org/0000-0001-9228-022X>

Sanghee Kim
<https://orcid.org/0009-0005-1107-7200>
Min Hwan Kwak
<https://orcid.org/0000-0003-0478-9507>
Ji Hyun Lee
<https://orcid.org/0000-0002-8582-5436>
Sang Yub Lee
<https://orcid.org/0000-0001-8529-8229>
Chang Hoon Oh
<https://orcid.org/0000-0002-1513-0612>
Sung Goo Park
<https://orcid.org/0009-0009-6508-8019>
Yun Hwa Roh
<https://orcid.org/0000-0002-8041-1621>
Yongsik Sim
<https://orcid.org/0000-0003-2711-2793>
Beomseok Sohn
<https://orcid.org/0000-0002-6765-8056>
Siyoun Sung
<https://orcid.org/0009-0002-2549-5129>
Min-Hyuk Yu
<https://orcid.org/0000-0003-2266-0707>
Seong Ho Park
<https://orcid.org/0000-0002-1257-8315>

Funding Statement

This study was supported by National Research Foundation of Korea (NRF), funded by the Korea government (MSIT) (No. RS-2025-00516874), and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2025-02309552).

Acknowledgments

The authors thank Kyunga Kim, PhD (Biomedical Statistics Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Korea), for providing statistical consultation and assistance with the data analysis for this study.

REFERENCES

1. Wang M, Le T, Leswick DA. Amplifying research: the potential for podcasts to boost Radiology journal article exposure. *Radiology* 2024;312:e233057
2. Kelly JM, Perseghin A, Dow AW, Trivedi SP, Rodman A, Berk J. Learning through listening: a scoping review of podcast use

- in medical education. *Acad Med* 2022;97:1079-1085
3. Tejani AS, Khosravi B, Savage CH, Moy L, Kahn CE, Yi PH. AI-generated podcast summaries of Radiology articles: analysis of content and quality. *Radiology* 2025;314:e243270
 4. OrthoMedia. Introduction of AI talks with bone & joint [accessed on November 4, 2025]. Available at: https://www.orthomedia.org.uk/Media/Introduction_Of_AI_Talks_With_Bone_Joint/522399f7-2ac1-4e2a-a18c-5d55e6c91e0a
 5. Radiological Society of North America. Episode 51: Radiology AI papers in a capsule series-episode 1 [accessed on November 4, 2025]. Available at: <https://rsnaradiologyai.libsyn.com/radiology-ai-papers-in-a-capsule-series-episode-1>
 6. BrJDermatol. We are delighted to announce the relaunch of the BJD Talks podcast! [accessed on November 4, 2025]. Available at: <https://x.com/BrJDermatol/status/1945854914694930828>
 7. Lee JH, Jang HY, Park SH. Academic journal podcast and generative artificial intelligence: introducing KJR SummaryCast. *Korean J Radiol* 2025;26:801-803
 8. Flanagan A, Kendall-Taylor J, Bibbins-Domingo K. Guidance for authors, peer reviewers, and editors on use of AI, language models, and chatbots. *JAMA* 2023;330:702-703
 9. TechCrunch. Google's NotebookLM now lets you guide AI-generated audio conversations, launches business pilot [accessed on January 12, 2026]. Available at: <https://techcrunch.com/2024/10/17/googles-notebooklm-now-lets-you-guide-ai-generated-audio-conversations-launches-business-pilot>
 10. Park SH, Suh CH, Lee JH, Tejani AS, You SC, Kahn CE, et al. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM): 2025 updates. *Korean J Radiol* 2025;26:1123-1132
 11. Park SH, Han K, Jang HY, Park JE, Lee JG, Kim DW, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* 2023;306:20-31
 12. Alarcón R, Blanca MJ. Development and psychometric properties of the questionnaire for assessing educational podcasts (QAEP). *Front Psychol* 2020;11:579454
 13. Committee on Publication Ethics. COPE position: authorship and AI tools [accessed on January 19, 2026]. Available at: <https://publicationethics.org/cope-position-statements/ai-author>
 14. International Committee of Medical Journal Editors. Recommendations [accessed on January 19, 2026]. Available at: <https://www.icmje.org/recommendations>
 15. Amano T, Ramírez-Castañeda V, Berdejo-Espinola V, Borokini I, Chowdhury S, Golivets M, et al. The manifold costs of being a non-native English speaker in science. *PLoS Biol* 2023;21:e3002184
 16. Koller D, Beam A, Manrai A, Ashley E, Liu X, Gichoya J, et al. Why we support and encourage the use of large language models in NEJM AI submissions. *NEJM AI* 2024;1:aIe2300128
 17. Google Workspace. Quickstart guide to Gemini and NotebookLM for education [accessed on November 14, 2025]. Available at: <https://support.google.com/a/answer/16350447>