Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

Using a Vision-Language Model to Generate Visual Abstracts for Radiology Journals

Taehee Lee, MD[*,1] • Seonhye Chae, MD*[*,1] • Seong Ho Park, MD, PhD*[2] • Charles E. Kahn, Jr, MD, MS*[3] • Seng Chan You, MD, PhD*[4] • Soon Ho Yoon, MD, PhD*[1]

* T.L. and S.C. contributed equally to this work.

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

See also the editorial by Chu and Syailendra in the October 2025 issue.

Radiology 2025; 316(3):e251458 • https://doi.org/10.1148/radiol.251458 • Content codes: AI ED

Visual abstracts (VAs) are increasingly used to improve the dissemination of scientific articles (1). Although some journals, such as *European Radiology*, require authors to submit VAs, others, such as *Radiology*, have them prepared by the editorial office (2). In both cases, creating VAs demands substantial time, visual design, and editorial effort. Recent vision-language models (VLMs), particularly GPT-4o (OpenAI), now enable end-to-end VA generation from full-text articles (3). However, the quality and suitability of VLM-generated VAs have not been evaluated in real-world publishing contexts. In this study, we generated VAs using GPT-4o for original research articles from *Radiology* and *European Radiology* and compared them with the published versions through blinded expert review.

Materials and Methods

This study did not involve human participants or patient data and was exempt from institutional review board approval.

To develop journal-specific prompts, all original research articles with VAs published online in March 2025 in *Radiology* (n = 19) and *European Radiology* (n = 36) were included. For testing, 75 articles with VAs published in January and February 2025 (38 from *Radiology* and 37 from *European Radiology*) were selected.

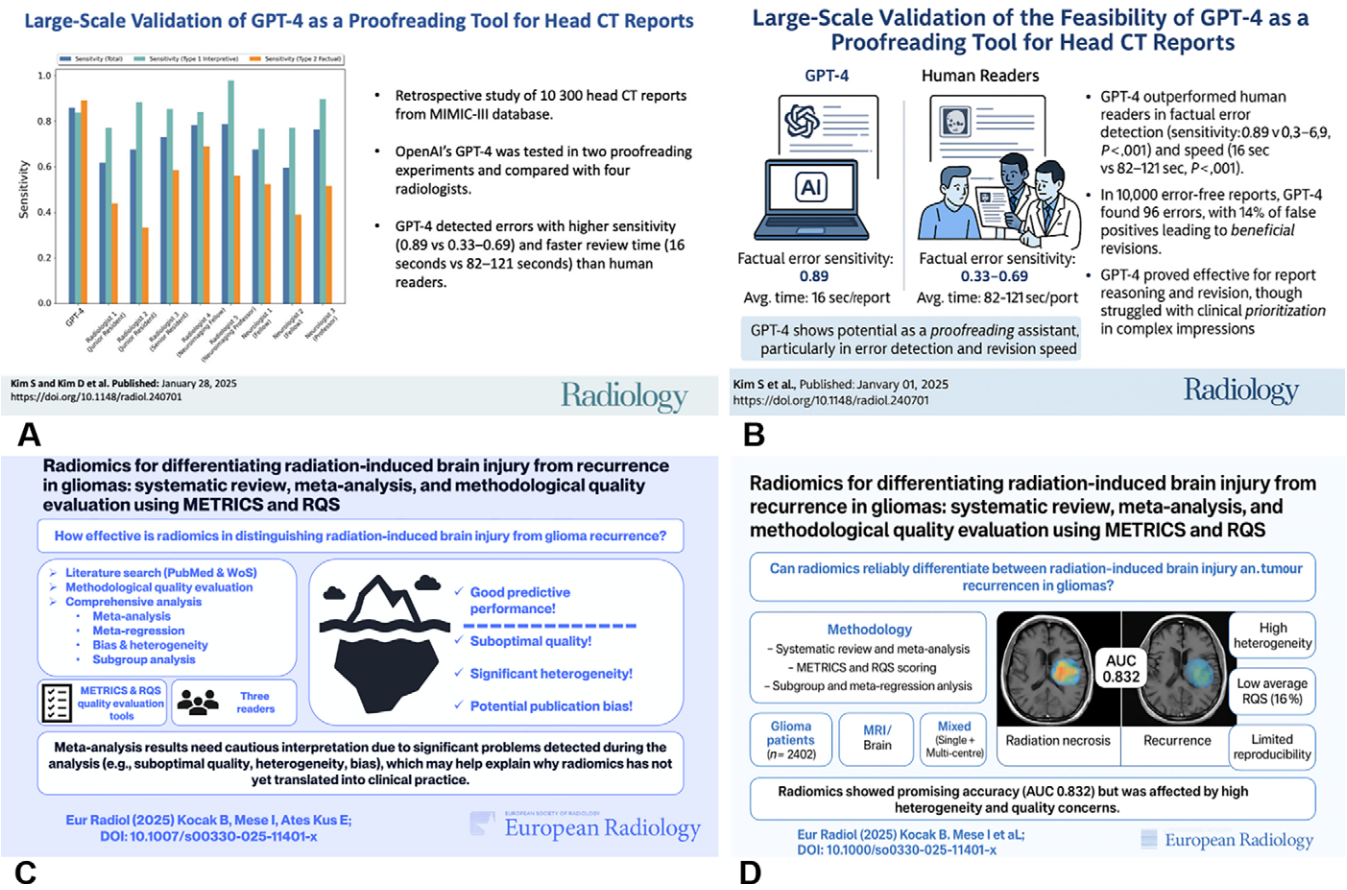
VAs were generated using GPT-4o via the ChatGPT web interface (chat.openai.com) from the Title, Abstract, Methods, and Results sections as input. Each VA was generated in a new session with memory disabled and no saving function to avoid context carryover. All generations occurred between April 4 and 6, 2025. Prompts tailored to each journal's layout and style, developed by a board-certified radiologist with prompt engineering expertise (T.L.), are available on GitHub *(https://github.com/lee8720/VLM_VA)*.

The design process followed three structured steps: extraction of key findings, confirmation of summary elements, and

Criteria Definitions and Expert Ratings for Original versus VLM-Generated Visual Abstracts

		Pooled			*Radiology*			*European Radiology*		
Criteria and Definition	Interreader Agreement	Original	VLM	*P* Value	Original	VLM	*P* Value	Original	VLM	*P* Value
Accuracy: how accurately the text and visual represent the article's key findings	0.04	4.56 ± 0.75	3.95 ± 1.06	<.001	4.49 ± 0.81	4.16 ± 0.88	<.001	4.64 ± 0.69	3.72 ± 1.19	<.001
Text clarity: how clearly and effectively the text conveys the main message	0.006	4.36 ± 0.80	4.23 ± 0.72	.03	4.47 ± 0.71	4.14 ± 0.82	<.001	4.24 ± 0.87	4.31 ± 0.72	.45
Visual quality: how well the visual supports the key message and maintains clarity	0.03	3.72 ± 0.98	3.43 ± 1.27	<.001	3.65 ± 1.00	3.51 ± 1.26	.22	3.80 ± 0.95	3.34 ± 1.29	<.001
Overall evaluation: overall quality considering accuracy, clarity, and visuals	0.02	3.92 ± 0.83	3.58 ± 1.06	<.001	3.93 ± 0.81	3.68 ± 1.00	.006	3.91 ± 0.84	3.47 ± 1.11	<.001
Preference: indicate which image you prefer	0.16	52 (157/300)	45 (135/300)	.22	47 (71/152)	49 (75/152)	.81	57 (86/148)	41 (60/148)	.04

Note.—Unless otherwise specified, data are mean scores with SDs. Interreader agreement was assessed using the intraclass correlation coefficient. Preference is presented as a percentage of selections for each visual abstract type (original vs vision-language model [VLM] generated), with raw counts shown in parentheses. Ties in preference occurred six times in *Radiology* and two times in *European Radiology*. *P* values were calculated using linear mixed-effects models for continuous variables and a nonparametric permutation-based sign test for preference data.



Examples of original and vision-language model (VLM)–generated visual abstracts (VAs) from **(A, B)** *Radiology* and **(C, D)** *European Radiology*. In the *Radiology* example, the **(A)** original VA received higher ratings for text clarity (mean score, 4.8 vs 4.5), whereas the **(B)** VLM-generated VA was rated higher for accuracy (mean score, 4.5 vs 5.0), visual quality (mean score, 3.3 vs 4.5), and overall evaluation (mean score, 4.0 vs 4.5). All four reviewers preferred the VLM-generated VA in this case. In the *European Radiology* example, the **(D)** VLM-generated VA received higher ratings for text clarity (mean score, 4.3 vs 4.8), visual quality (mean score, 3.8 vs 4.0), and overall evaluation (mean score, 3.5 vs 4.3); both VAs received the same score for accuracy. Again, all four reviewers preferred the VLM-generated VA. The original VA in **A** is reprinted, with permission, from reference 4. The original VA in **C** is reprinted, under a CC BY 4.0 license, from reference 5 *(https://creativecommons.org/licenses/by/4.0/)*. GPT-4o (OpenAI) was used to generate VAs in **B** and **D**.

layout-specific graphical rendering. To ensure adherence to journal format, prompts included paired layout templates and example images as style references.

Four expert reviewers, all editors or editorial board members of leading radiology and medical journals (S.H.P., C.E.K., S.C.Y., and S.H.Y.), evaluated each randomized, blinded VA pair. Each image was rated on a five-point Likert scale (5 indicating the best) for accuracy, text clarity, visual quality, and overall evaluation. Reviewers also indicated their preference from each pair. Interrater agreement was assessed using the intraclass correlation coefficient. Likert scores were analyzed using linear mixed models. Preference data were analyzed using a nonparametric sign test with permutation resampling.

Results

Interrater agreement was low across all evaluation criteria, with intraclass correlation coefficient values below 0.16. Original VAs received higher ratings than VLM-generated VAs across all four criteria, though the score differences were approximately 0.5 points or less, with means scores of 4.56 ± 0.75 (SD) versus 3.95 ± 1.06, respectively, for accuracy ($P < .001$), 4.36 ± 0.80 versus 4.23 ± 0.72 for text clarity ($P = .029$), 3.72 ± 0.93 versus 3.43 ± 1.27 for visual quality ($P < .001$), and 3.92 ± 0.83 versus 3.58 ± 1.06 for overall evaluation ($P < .001$) (Table, Figure). There was no evidence of a difference in preference between original and VLM-generated VAs (52% [157 of 300 evaluations] vs 45% [135 of 300], including eight ties; $P = .22$).

Discussion

Compared with state-of-the-art VLM-generated VAs, human-created VAs demonstrated modest superiority across evaluation criteria except for accuracy. Notably, no statistically significant preference difference emerged between the two approaches, particularly within the *Radiology* subgroup (47% vs 49%; $P = .81$). However, as no a priori power calculation was performed, smaller differences may have remained undetected. These findings suggest that despite current VLMs exhibiting limitations in VA production, primarily due to hallucination artifacts, they demonstrate potential in addressing the time-consuming challenge of VA creation faced by authors and editorial offices.

Reviewers noted that VLM-generated VAs often conveyed core concepts clearly and succinctly, whereas many original VAs simply reused manuscript figures that were not specifically designed for visual abstraction. However, because the VLM could not reuse any original images due to copyright and platform restrictions, this may have affected the comparison.

Interestingly, some reviewers, including those less familiar with the domain, found the VLM-generated versions more accessible and informative.

Observed limitations of VLM-generated outputs included typographic artifacts, occasional labeling errors, and synthetic imagery resulting from the inability to reuse original figures. Although these challenges reflect broader issues in layout-aware generation and visual fidelity, they also point to a clear opportunity for human input—particularly in reviewing and refining artificial intelligence–generated visuals for publication.

Rather than replacing human effort, VLMs can serve as effective first-draft tools that, when combined with editorial input via human–artificial intelligence codevelopment, may improve the efficiency and quality of VA production.

Although this study did not evaluate workflow efficiency, future research could explore how VLMs may support editorial office–driven VA production, such as in *Radiology*. Evaluating how these tools affect overall editorial processes—beyond time savings—may provide insight into their practical integration into publication workflows.

Deputy Editor: Kathryn Fowler

Author affiliations:

[1] Department of Radiology, Seoul National University Hospital and College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea

[2] Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

[3] Department of Radiology, University of Pennsylvania, Philadelphia, Pa

[4] Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea

Received May 13, 2025; revision requested June 13; final revision received June 30; accepted August 12.

Address correspondence to: S.H.Y. (email: yshoka@gmail.com).

Funding: Supported by the Naver Digital Bio Innovation Research Fund, funded by Naver Corporation (grant no. 3720230020). However, the funder had no role in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; and in the decision to submit the article for publication.

Author contributions: Guarantor of integrity of entire study, **S.H.Y.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **T.L., S.H.Y.**; clinical studies, **T.L.**; experimental studies, **T.L., S.C., S.H.P., C.E.K., S.H.Y.**; statistical analysis, **T.L., S.C.Y.**; and manuscript editing, all authors

Disclosures of conflicts of interest: **T.L.** Assistant editor of the *American Journal of Roentgenology*. **S.C.** No relevant relationships. **S.H.P.** Member of the *Radiology* editorial board. **C.E.K.** Salary support from RSNA paid to employer for service as Editor of *Radiology: Artificial Intelligence*; member of the *Radiology* editorial board. **S.C.Y.** Grants from Daiichi-Sankyo; associate editor of *Journal of the American College of Cardiology* (JACC); chief executive officer of PHI Digital Healthcare. **S.H.Y.** Stock or stock options in MEDICAL IP; associate editorial board member of *Investigative Radiology*.

References

1. Brook OR, Vernuccio F, Nicola R, Cannella R, Altinmakas E. Visual abstract for Abdominal Radiology: what it is, why we need it and how to make it. Abdom Radiol (NY) 2021;46(6):2403–2406.
2. Kelly BS. Embracing graphical abstracts in European Radiology. Eur Radiol 2025. 10.1007/s00330-025-11555-8. Published online April 3, 2025.
3. OpenAI. Introducing 4o Image Generation. https://openai.com/index/introducing-4o-image-generation/. Accessed April 1, 2025.
4. Kim S, Kim D, Shin HJ, et al. Large-Scale Validation of the Feasibility of GPT-4 as a Proofreading Tool for Head CT Reports. Radiology. 2025;314(1):e240701.
5. Kocak B, Mese I, Ates Kus E. Radiomics for Differentiating Radiation-induced Brain Injury from Recurrence in Gliomas: Systematic Review, Meta-analysis, and Methodological Quality Evaluation using METRICS and RQS. Eur Radiol. 2025;35(8):4490-4505.