Text-Deficient Multimodal Stroke Segmentation with Lesion-Grounded Self-Retrieval-Augmented Generation

Heeseong Eum*[1], Junhyeok Lee[1], Joon Jang[2], Han Jang[3], Songsoo Kim[4], and Kyu Sung Choi†[4,5,6]

[1] Interdisciplinary Program in Cancer Biology, Seoul National University College of Medicine, Seoul, South Korea
[2] Dept. of Biomedical Sciences, Seoul National University, Seoul, South Korea
[3] Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, South Korea
[4] Dept. of Radiology, Seoul National University Hospital, Seoul, South Korea
[5] Dept. of Radiology, Seoul National University College of Medicine, Seoul, South Korea
[6] Healthcare AI Research Institute, Seoul National University Hospital, Seoul, South Korea
seong6466@snu.ac.kr; ent1127@snu.ac.kr

Abstract. Acute ischemic stroke requires rapid delineation of infarct core on diffusion-weighted imaging (DWI) and apparent diffusion coefficient (ADC) to guide time-critical decisions. Clinical urgency yields preliminary reporting as an early workflow summary centered on lesion location and extent, which enables report-conditioned multimodal segmentation. Paradoxically, the same urgency also makes preliminary reports frequently missing or inconsistently archived in retrospective cohorts, creating a key bottleneck for multimodal stroke segmentation. To address this gap, we present (i) Lesion-Grounded Self-Retrieval-Augmented Generation (LeG-RAG), a training pipeline that retrieves lesion-concordant reports using atlas-normalized lesion masks and synthesizes missing reports with a multimodal large language model without additional training, and (ii) LLMSwin, a report-conditioned multimodal segmentation architecture for brain MRI that extracts text features by combining the report with learnable prompts and feeding them into a frozen Llama-3-8B-Instruct model, and fuses these representations with multi-scale features from a SwinUNETR 3D backbone via two-way transformers. Experiments on a multi-institutional 3D stroke MRI dataset show that LeG-RAG achieves superior retrieval performance compared with state-of-the-art image-embedding methods across text-deficiency regimes, while remaining diagnostically consistent under ASPECTS-based reliability checks. LLMSwin achieves state-of-the-art segmentation across text-availability settings, outperforming image-only and prior multimodal segmentation baselines. Code is available at `https://github.com/HeeseongEom/LeG-RAG`.

* First author. † Corresponding author.

Keywords: Multimodal Segmentation · Stroke MRI · Retrieval-Augmented Generation · Synthesized Report · Large Multimodal Model

1 Introduction

Acute ischemic stroke management hinges on rapid localization of infarct core on DWI/ADC, where minutes can influence treatment eligibility and outcome [3, 11]. In routine neuroradiology, early communication often takes the form of preliminary reporting: a short impression emphasizing *where* the lesion is (laterality, standardized regions, vascular territories) and *how extensive* it appears [6,8]. However, preliminary texts are often absent or heterogeneously stored in retrospective imaging cohorts, despite availability of images and lesion masks.

We mitigate this mismatch by synthesizing missing reports during training via lesion-grounded retrieval. Since the multimodal LLM can ingest the *query image* at synthesis time, the information most likely to require external grounding is fine-grained spatial semantics (location/extent), which are explicitly encoded in lesion masks and dominate preliminary-style descriptions. After spatial normalization to a common atlas space (e.g., MNI), lesion masks become directly comparable across patients and align with atlas-defined regions used in clinical reasoning. We therefore propose *LeG-RAG*, which retrieves report exemplars from the same institutional archive using a coarse-to-fine, mask-driven strategy and generates a concise *synthesized report* via in-context learning [2, 7, 12]. We evaluate clinical reliability using the *Alberta Stroke Program Early CT Score (ASPECTS)*, a standardized regional scoring scheme [1, 10].

In parallel, we introduce a dedicated report-conditioned multimodal segmentation model tailored to volumetric brain MRI. Inspired by LLM-based multimodal segmentation for target contouring [9], we extend the paradigm to stroke MRI and integrate it with a strong 3D backbone. Our *LLMSwin* encodes reports with a frozen pretrained LLM using learnable prompts and fuses the resulting representation with multi-scale SwinUNETR features [4] through two-way transformers at each stage. We also instantiate a UNETR-based variant (*LLMUNETR*) [5] using the same report pathway; LLMSwin consistently performs best, aligning with the empirical suitability of SwinUNETR-style representations for brain MRI [4].

In summary, we propose (i) **LeG-RAG**, a lesion-grounded retrieval-and-synthesis framework for text-deficient stroke training; (ii) **LLMSwin**, a report-conditioned volumetric multimodal segmentation architecture; and (iii) comprehensive evaluation showing improved synthesized-report quality and robust segmentation under text deficiency.

2 Method

2.1 Problem Setup

We study training under a text availability ratio p (in %), defined as the fraction of cases with available preliminary reports. The dataset is partitioned into a

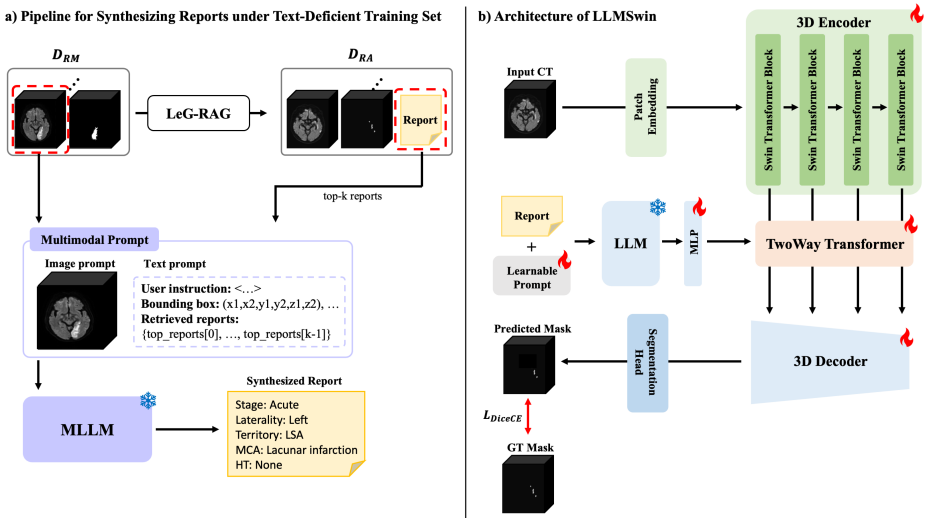

Fig. 1. (a) LeG-RAG overview: for each report-missing case in D_{RM}, retrieve lesion-concordant reports from D_{RA} in atlas space and synthesize a preliminary-style report for multimodal training. (b) Overview of the proposed **LLMSwin** report-conditioned volumetric segmentation architecture.

report-available subset

$$D_{RA} = \{(v_j, r_j, m_j)\}_{j=1}^{R}, \tag{1}$$

and a report-missing subset

$$D_{RM} = \{(v_i, m_i)\}_{i=1}^{S}, \tag{2}$$

where $N = R + S$ and $p = \frac{R}{N} \times 100$. Each v is a 3D MRI input formed by paired DWI and ADC volumes (two channels in $\mathbb{R}^{H\times W\times D\times 2}$), each $m \in \{0,1\}^{H\times W\times D}$ is a binary infarct mask, and each r is a short preliminary report. All volumes and masks are spatially normalized to a common atlas space (e.g., MNI). Our goal is to generate a synthesized report $\hat{r}_i$ for each $(v_i, m_i) \in D_{RM}$, yielding an augmented multimodal training set

$$D' = D_{RA} \cup \{(v_i, \hat{r}_i, m_i)\}_{i=1}^{S}, \tag{3}$$

which is then used to train report-conditioned multimodal segmentation models.

2.2 Lesion-Grounded Self-Retrieval-Augmented Generation

Spatial Key Retriever (SKR) For each query $(v_i, m_i) \in D_{RM}$, SKR computes an axial lesion profile $S_i \in \mathbb{R}^D$ by counting lesion voxels per axial slice and normalizing to unit length. For each candidate $(v_j, r_j, m_j) \in D_{RA}$, we compute

S_j likewise. SKR retains candidates whose cosine similarity exceeds a threshold δ:

$$S_i = SK(m_i), \quad S_j = SK(m_j), \quad FD_{RA}(i) = \{(v_j, r_j, m_j) \in D_{RA} \mid \cos(S_i, S_j) > \delta\}. \tag{4}$$

We use $\delta = 0$ by default.

Clinical Context Retriever (CCR) CCR ranks candidates in $FD_{RA}(i)$ using a clinical context score combining overlap and boundary consistency of lesion masks. Let $\mathrm{DSC}(\cdot,\cdot)$ be Dice similarity and $\mathrm{HD95}(\cdot,\cdot)$ the 95th percentile Hausdorff distance. We define:

$$\mathrm{CCS}(m_i, m_j) = \lambda\, \mathrm{DSC}(m_i, m_j) + (1-\lambda)\, \frac{1}{1 + \mathrm{HD95}(m_i, m_j)}, \tag{5}$$

with $\lambda = \frac{1}{2}$. The top-k exemplars are:

$$\mathcal{T}_i = \text{Top-}k_{(v_j, r_j, m_j) \in FD_{RA}(i)} \big[\, \mathrm{CCS}(m_i, m_j) \,\big]. \tag{6}$$

Synthesized Report Generation Given $\mathcal{T}_i$, we provide the MLLM with the query image v_i and the retrieved reports $\{r_j\}$ as in-context demonstrations, and generate a concise synthesized report:

$$\hat{r}_i = \mathrm{MLLM}\big(\text{instruction};\ v_i, m_i, \{(m_j, r_j)\}_{(v_j, r_j, m_j) \in \mathcal{T}_i}\big). \tag{7}$$

Applying this to all $(v_i, m_i) \in D_{RM}$ yields D' for multimodal segmentation training.

2.3 LLMSwin: Report-Conditioned Multimodal Segmentation

LLMSwin integrates report semantics into a SwinUNETR-style volumetric backbone [4] using a frozen pretrained LLM and two-way transformer fusion. Let the Swin encoder produce multi-scale feature maps $\{F_\ell\}_{\ell=1}^{L}$, where $F_\ell \in \mathbb{R}^{H_\ell \times W_\ell \times D_\ell \times C_\ell}$. For a report r, we prepend learnable prompt tokens P and a special token $\langle \mathrm{seg} \rangle$, and feed the token sequence into a frozen pretrained LLM:

$$H = \mathrm{LLM}\big([P;\ \langle \mathrm{seg} \rangle;\ \mathrm{Tok}(r)]\big), \qquad h_{\mathrm{seg}} = H_{\langle \mathrm{seg} \rangle} \in \mathbb{R}^{d_L}. \tag{8}$$

For each encoder scale ℓ, we project h_{seg} to match the channel dimension:

$$z_\ell = \mathrm{MLP}_\ell(h_{\mathrm{seg}}) \in \mathbb{R}^{C_\ell}. \tag{9}$$

We fuse z_ℓ with visual tokens of F_ℓ using a two-way transformer TWT_ℓ. Let $\phi(\cdot)$ flatten spatial dimensions into tokens:

$$(\tilde{X}_\ell, \tilde{z}_\ell) = \mathrm{TWT}_\ell\big(\phi(F_\ell),\ z_\ell\big), \qquad \tilde{F}_\ell = \phi^{-1}(\tilde{X}_\ell). \tag{10}$$

Table 1. Stroke segmentation DSC on Test A/B under text-deficient training (text availability p). Img: image-only. MM-P: multimodal with available real reports only. MM-SR: multimodal with LeG-RAG synthesized reports. MM-Full: multimodal with all real reports (upper bound). Best results among Img, MM-P, and MM-SR at each text availability are in **bold**.

Model	Text avail.	Test A				Test B			
		Img	MM-P	**MM-SR (ours)**	MM-Full	Img	MM-P	**MM-SR (ours)**	MM-Full
LLMSeg	20%	0.708	0.684	**0.719**	0.729	0.615	0.595	**0.649**	0.645
	40%	0.708	0.700	**0.722**	0.729	0.615	0.597	**0.659**	0.645
	60%	0.708	0.715	**0.725**	0.729	0.615	0.631	**0.645**	0.645
	80%	0.708	0.719	**0.723**	0.729	0.615	0.612	**0.652**	0.645
LLMUNETR	20%	0.628	0.652	**0.664**	0.668	0.551	0.591	**0.601**	0.612
	40%	0.628	0.671	**0.689**	0.668	0.551	0.568	**0.601**	0.612
	60%	0.628	**0.676**	0.674	0.668	0.551	**0.617**	0.586	0.612
	80%	0.628	0.670	**0.696**	0.668	0.551	0.564	**0.624**	0.612
LLMSwin	20%	0.705	0.722	**0.755**	0.755	0.623	0.589	**0.685**	0.671
	40%	0.705	0.739	**0.755**	0.755	0.623	0.652	**0.661**	0.671
	60%	0.705	0.746	**0.752**	0.755	0.623	0.666	**0.664**	0.671
	80%	0.705	0.743	**0.749**	0.755	0.623	0.667	**0.673**	0.671

The fused features $\{\tilde{F}_\ell\}$ are forwarded to a 3D decoder with skip connections to predict the lesion mask $\hat{m}$:

$$\hat{m} = \mathrm{Dec}(\{\tilde{F}_\ell\}_{\ell=1}^{L}). \tag{11}$$

Training minimizes a combined Dice and cross-entropy loss:

$$\mathcal{L}_{\mathrm{DiceCE}} = \mathcal{L}_{\mathrm{Dice}}(\hat{m}, m) + \mathcal{L}_{\mathrm{CE}}(\hat{m}, m). \tag{12}$$

Only prompts, projection MLPs, fusion modules, and the segmentation backbone/decoder are optimized; the pretrained LLM remains frozen.

3 Experiments and Results

3.1 Dataset and Evalutaion

Dataset and text availability protocol We evaluate on a multi-institutional 3D stroke MRI cohort from three hospitals (anonymized). The training split contains 441 cases, and evaluation is performed on two held-out test sets (Test A: 112 cases; Test B: 151 cases). All images underwent skull stripping, resampling to 2 mm isotropic resolution, and Z-score normalization, followed by spatial normalization to a common atlas space. We simulate text-deficient training by varying text availability $p \in \{20, 40, 60, 80\}\%$ on the training set; for each p, D_{RA} contains the text-available subset and D_{RM} contains the remaining report-missing subset.

Table 2. Evaluation of synthesized report under varying text availability p (aggregated F1 and accuracy) across embedding-based retrieval baselines. Best/second-best are in **bold**/underline.

Model	Text availability							
	20%		40%		60%		80%	
	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑
Rad_DenseNet121	0.500	0.487	0.507	0.486	0.475	0.458	0.492	0.473
OpenCLIP	0.526	0.514	0.533	0.517	<u>0.537</u>	<u>0.522</u>	0.549	0.530
Rad_InceptionV3	<u>0.527</u>	0.509	0.534	0.517	0.528	0.516	0.535	0.523
BiomedCLIP	0.526	0.507	<u>0.535</u>	0.514	0.523	0.507	<u>0.560</u>	<u>0.543</u>
UNI-2	0.524	<u>0.519</u>	0.524	<u>0.520</u>	0.517	0.512	0.518	0.509
LeG-RAG (ours)	**0.715**	**0.671**	**0.733**	**0.713**	**0.754**	**0.728**	**0.750**	**0.711**

Evaluation metrics We evaluate segmentation using DSC. For synthesized report quality, we extract clinical attributes from real and synthesized reports using the same structured pipeline and report aggregated F1 and accuracy. For diagnostic consistency, we compute ASPECTS region-involvement agreement (F1/Acc) and total-score MAE from atlas-space lesion involvement [1, 10].

3.2 Quantitative Analysis

Segmentation under text deficiency Table 1 shows that MM-SR improves over MM-P, particularly at low p, suggesting that LeG-RAG synthesized reports provide useful supervision when real reports are scarce. LLMSwin achieves the strongest performance across all p, outperforming image-only training and multimodal baselines, consistent with the strong volumetric representations of SwinUNETR-style backbones [4].

ASPECTS reliability and diagnostic consistency Table 4 summarizes ASPECTS-based agreement derived from atlas-space lesion involvement [1, 10]. Across models and test sets, MM-SR improves or matches region-wise agreement while reducing MAE relative to text-deficient baselines, supporting diagnostic consistency under synthesized-report-driven training.

Synthesized report quality Table 2 shows that LeG-RAG retrieval yields higher attribute agreement than image-embedding retrieval baselines across all p, consistent with the benefit of explicit spatial grounding over generic embedding similarity [2, 7]. Under fixed LeG-RAG retrieval, Table 3 indicates that synthesis quality depends on the chosen MLLM, while retrieval provides a strong, model-agnostic foundation.

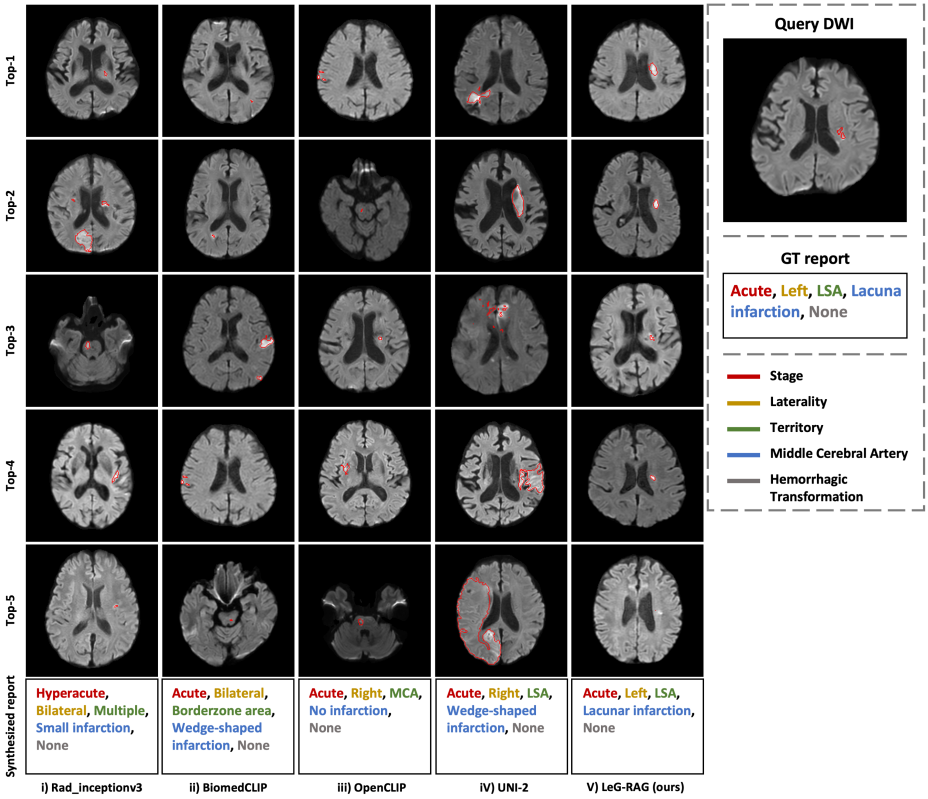

Fig. 2. Example retrievals and synthesized reports. LeG-RAG retrieves lesion-concordant exemplars in atlas space, yielding reports that better match laterality and territory-related attributes than embedding-based retrieval.

Table 3. MLLM comparison for report synthesis under fixed **LeG-RAG** retrieval (aggregated F1 and accuracy across text availability p). Best/second-best are in **bold**/underline.

Model	**Text availability**							
	20%		40%		60%		80%	
	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑
Qwen2.5	0.691	0.795	0.725	0.808	0.739	0.815	0.750	0.825
Qwen2.5 (multi-slice)	0.693	**0.816**	0.725	**0.823**	0.736	**0.815**	0.749	**0.834**
VideoLLaVA	0.670	0.680	0.729	0.741	0.742	0.756	**0.774**	0.789
GPT-o1	0.730	0.763	0.749	0.801	0.770	0.800	0.773	0.806
LLaVA-OneVision	0.741	0.757	0.756	0.773	0.776	0.798	0.760	0.783
LLaVA-OneVision (multi-slice)	**0.749**	0.767	**0.765**	0.783	**0.777**	0.800	0.763	0.780

Table 4. ASPECTS-based diagnostic consistency on Test A/B (region involvement F1/Acc, total-score MAE). Img: image-only. MM-P: multimodal with available real reports only. **MM-SR**: multimodal with LeG-RAG synthesized reports. MM-Full: multimodal with all real reports (upper bound). Bold: best among Img/MM-P/MM-SR.

Model	Setting	Test A			Test B		
		F1↑	Acc↑	MAE↓	F1↑	Acc↑	MAE↓
LLMSeg	Img	0.879	0.929	0.411	0.909	0.919	0.748
	MM-P	0.916	0.951	0.393	0.912	0.920	0.737
	MM-SR (ours)	**0.944**	**0.967**	**0.315**	**0.934**	**0.939**	**0.657**
	MM-Full	0.954	0.973	0.339	0.927	0.933	0.726
LLMUNETR	Img	0.829	0.893	0.509	0.883	0.896	0.807
	MM-P	0.858	0.908	0.496	0.885	0.894	0.828
	MM-SR (ours)	**0.862**	**0.911**	**0.451**	**0.901**	**0.909**	**0.817**
	MM-Full	0.857	0.911	0.491	0.893	0.904	0.852
LLMSwin	Img	0.886	0.929	0.393	0.927	0.933	0.711
	MM-P	0.879	0.922	0.344	0.923	0.930	0.654
	MM-SR (ours)	**0.910**	**0.942**	**0.283**	**0.937**	**0.941**	**0.604**
	MM-Full	0.907	0.938	0.295	0.929	0.933	0.622

Table 5. Ablation of CCS components in **LeG-RAG** retrieval (top-5 average attribute F1 across text availability p).

Setting	Text availability			
	20%	40%	60%	80%
LeG-RAG + DSC only	0.547	0.558	0.616	0.635
LeG-RAG + HD95 only	<u>0.633</u>	<u>0.660</u>	<u>0.686</u>	<u>0.702</u>
LeG-RAG + CCS (DSC+HD95)	**0.635**	**0.663**	**0.694**	**0.713**

3.3 Qualitative Analysis

Figure 2 provides representative examples, where LeG-RAG selects exemplars with closer lesion distribution in atlas space and yields synthesized reports that better reflect laterality and territory-relevant descriptions.

3.4 Ablation Analysis

Table 5 shows that using only DSC degrades retrieval quality, while the HD95-derived term is more robust; the combined CCS yields the best performance, supporting complementary overlap and boundary cues.

4 Conclusion

We presented **LeG-RAG**, a lesion-grounded retrieval and report-synthesis pipeline for text-deficient stroke training that leverages atlas-aligned lesion masks to retrieve spatially relevant exemplars and generate preliminary-style reports via in-context learning [2,7,12]. We also introduced **LLMSwin**, which fuses prompt-conditioned report representations with multi-scale SwinUNETR volumetric features [4] via two-way transformer fusion. Across multi-institutional stroke MRI, LeG-RAG improves attribute agreement and ASPECTS-based consistency [1, 10], while LLMSwin achieves state-of-the-art segmentation across text availability settings.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25479661) (K.S.C.); the SNUH Research Fund (No. 04-2024-0600; No. 04-2025-2060) (K.S.C.); and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) grant funded by the Ministry of HealthWelfare (No. RS-2024-00439549) (K.S.C.).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Barber, P., Demchuk, A., Zhang, J., Buchan, A.: Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. Lancet **355**, 1670–1674 (2000)
2. Chen, W., Hu, H., Chen, X., Verga, P., Cohen, W.: Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In: EMNLP 2022. pp. 5558–5570. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.emnlp-main.375
3. Goyal, M., Menon, B., van Zwam, W., Dippel, D., Mitchell, P., Demchuk, A., Dávalos, A., Majoie, C., van der Lugt, A., de Miquel, M., Donnan, G., Roos, Y., Bonafe, A., Jahan, R., Diener, H.C., van den Berg, L., Levy, E., Berkhemer, O., Pereira, V., Rempel, J., Millán, M., Davis, S., Roy, D., Thornton, J., Román, L., Ribó, M., Beumer, D., Stouch, B., Brown, S., Campbell, B., van Oostenbrugge, R., Saver, J., Hill, M., Jovin, T.: Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet **387**, 1723–1731 (2016)
4. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: BrainLes Workshop, MICCAI. pp. 272–284. LNCS, Springer (2022)
5. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: WACV 2022. pp. 574–584 (2022)

6. Issa, G., Taslakian, B., Itani, M., Hitti, E., Batley, N., Saliba, M., El-Merhi, F.: The discrepancy rate between preliminary and official reports of emergency radiology studies: a performance indicator and quality improvement method. Acta Radiologica **56**(5), 598–604 (2015). https://doi.org/10.1177/0284185114532922
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: NeurIPS 2020. pp. 9459–9474 (2020)
8. Mates, J., Branstetter, B.F., Morgan, M.B., Lionetti, D.M., Chang, P.J.: 'wet reads' in the age of pacs: Technical and workflow considerations for a preliminary report system. Journal of Digital Imaging **20**(3), 296–306 (2007). https://doi.org/10.1007/s10278-006-1049-y
9. Oh, Y., Park, S., Byun, H., Cho, Y., Lee, I., Kim, J., Ye, J.: Llm-driven multimodal target volume contouring in radiation oncology. Nat. Commun. **15**, 9186 (2024)
10. Pexman, J., Barber, P., Hill, M., Sevick, R., Demchuk, A., Hudon, M., Hu, W., Buchan, A.: Use of the alberta stroke program early ct score (aspects) for assessing ct scans in patients with acute stroke. AJNR Am. J. Neuroradiol. **22**, 1534–1542 (2001)
11. Saver, J.L.: Time is brain—quantified. Stroke **37**(1), 263–266 (2006). https://doi.org/10.1161/01.STR.0000196957.55928.ab
12. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. NeurIPS **35**, 24824–24837 (2022)