Evaluating Generative AI for Medical Image Captioning: A Benchmark Study on Radiology Images

Heitor Mendes Pereira[a], Pedro Didier Maranhão[a], Vitoria de Araújo Xavier[a], Tsang Ing Ren[a], Anoushka Duggal[b], Alba G. Seco de Herrera[c], and Vahid Abolghasemi[d]

[a]Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, Brazil
[b]Thapar Institute of Engineering & Technology, Patiala, Punjab, India
[c]UNED, Spain
[d]School of Computer Science & Electronic Engineering, University of Essex, CO4 3SQ, Colchester, U.K.

ABSTRACT

The generation of clinically accurate and contextually rich captions for radiology images is a critical task in medical Artificial Intelligence (AI), with applications in education, documentation, and decision support. In this study, we benchmark the performance of leading generative AI models—including OpenAI's GPT 4o, Google's Gemini 2.5 pro, Anthropic's Claude 4.5 Sonnet and Meta's LLaMA 4—on the recently released ROCOv2 (Radiology Objects in COntext Version 2) dataset. ROCOv2 offers a large-scale, multimodal resource of radiology images paired with expert-generated captions, enabling robust evaluation of vision-language models in the medical domain. We assess models under zero-shot and few-shot prompting conditions, and evaluate outputs using automated metrics (BLEU, BERTScore, ROUGE, METEOR and CIDEr). Our analysis highlights the capabilities and limitations of current generative models in understanding and describing complex radiological content, discussing the potential for integrating these models into clinical workflows. This work provides a comprehensive evaluation of generative AI for medical image captioning and offers insights into future directions for improving reliability and clinical relevance in multimodal medical AI systems.

Keywords: Generative AI, Multimodal Foundation Models (MFMs), Medical Image Captioning, Radiology Imaging, Computer-Aided Diagnosis, ImageCLEFmed, ROCOv2 dataset, Benchmarking

1. INTRODUCTION

The purpose of this study is to evaluate the performance of state-of-the-art generative Artificial Intelligence (AI) models in the task of medical image captioning,[1,2] with a specific focus on the complex domain of radiology.[3] This work is motivated by the growing interest in multimodal AI systems capable of interpreting and describing medical images in natural language, a capability with significant implications for clinical documentation, education, and diagnostic decision support.[4–6]

This study is situated within the long-standing framework of the Conference and Labs of the Evaluation Forum (CLEF),[7] specifically under the auspices of the ImageCLEF lab, a dedicated initiative for the cross-language evaluation of image retrieval and classification. The year 2026 marks a significant milestone: the 10th anniversary of the ImageCLEFmed Caption task.[8] Since its inception, this task has served as an international benchmark for automated clinical image interpretation, tracking the field's evolution from early concept detection to complex narrative generation. While the first decade of the ImageCLEFmed Caption task primarily focused on traditional machine learning and early deep learning architectures, the present work commemorates this decennial by exploring the next frontier: Multimodal Foundation Models (MFMs). We assess the capabilities of MFMs such as OpenAI's GPT-4 and Google's Gemini using the ROCOv2 dataset,[9] a recently released, expert-annotated dataset that offers a richer, more diverse set of radiology images and captions than its predecessors.

Model outputs are compared under different prompting conditions and evaluated using automated metrics (BLEU, BERTScore, ROUGE, METEOR and CIDEr). Specifically, this study aims to:

Further author information: (Send correspondence to A.G.S.H.)
A.G.S.H.: E-mail: alba.garcia@lsi.uned.es, Telephone:+34 91 398 8736

- Benchmark the current capabilities of generative AI in producing clinically accurate and contextually appropriate image captions;
- Explore the impact of few-shot prompting on the quality and reliability of generated text;
- Identify specific failure modes and linguistic limitations of MFMs when applied to specialized radiological content;
- Provide evidence-based insights into the clinical applicability of these models and their potential integration into real-world workflows.

Ultimately, this work contributes to the development of reliable, explainable, and clinically relevant multimodal AI systems and offers a forward-looking perspective on the role of generative models as medical imaging enters a new era of AI-driven interpretation.

The remainder of this paper is organized as follows. Section 2 reviews the evolution of medical image captioning, contrasting early deep learning architectures with the recent emergence of MFMs. Section 3 details our experimental framework, including a description of the ROCOv2 dataset, the generative AI models evaluated (GPT-4, Gemini 2.5 Pro, Claude 4.5, and LLama 4) in medical imaging. In Section 4, the qualitative and quantitative results are presented. Section 5, provides the discussion on the achieved results, and Section 6 concludes the paper.

2. RELATED WORK

Medical image captioning aims to automatically generate descriptive and clinically informative text from medical images, supporting tasks such as report drafting, documentation, and clinical decision support. Earlier work in this domain largely adapted classical deep learning approaches, combining convolutional encoders with recurrent language decoders to generate radiology captions and summaries. A seminal example combines a Show-Attend-Tell architecture with a generative pretrained transformer, demonstrating the feasibility of integrating visual features with language models to describe chest X-ray findings on public datasets such as MIMIC-CXR and Open-I.[10]

Recent advances in multimodal vision-language models (VLMs) have shifted the field toward larger, more flexible architectures that leverage joint representations of images and text. Vision-language pretraining strategies have been shown to benefit a range of medical tasks, including report generation, classification, and cross-modal retrieval, by learning unified feature spaces that generalize across downstream applications.[11] Additionally, transformer-based generative models conditioned on visual input have demonstrated improved ability to capture semantic relationships between image content and descriptive text.

Within the medical image captioning literature, specialized adaptations of large VLMs tailored to clinical contexts have shown promising results. For example, BLIP-based models have been applied to caption generation tasks in medical challenges such as ImageCLEFmedical, where methods exploiting pretrained text and image backbones perform competitively on standard metrics.[12] More recently, DualPrompt-MedCap introduced modality-aware and question-guided prompt strategies to enhance caption generation performance on medically relevant datasets, achieving substantial gains over baseline VLM captioners by explicitly encoding clinical context.[13] Similarly, Tran et al. proposed the VTG-Transformer for chest X-ray captioning, which integrates detection of clinical signs and their relationships into caption generation, improving both accuracy and interpretability of captions on benchmark datasets.[14] In ophthalmology, GCS-M3VLT uses guided context self-attention to combine multimodal retinal image features and textual information, demonstrating improved caption quality on retinal image datasets.[15]

Explainability and interpretability have gained increased attention in medical image captioning as a means to build trust and transparency into AI systems. Kamal et al. introduced an explainable captioning framework that combines vision transformers with explainable AI techniques such as Layer-wise Relevance Propagation (LRP) and Local Interpretable Model-agnostic Explanations (LIME) to highlight image regions influencing caption generation, making the model's reasoning more accessible to clinicians.[16] Such approaches aim to bridge the gap

between high caption performance and clinical reliability by providing visual and textual explanations of model outputs.

Benchmarking and systematic evaluation of VLMs for radiology captioning tasks remain active areas of research. A recent empirical study evaluated a suite of vision-language models with low-rank adaptation strategies on the ROCOv2 dataset across multiple imaging modalities, providing insights into parameter-efficient adaptation and performance trade-offs in practical captioning scenarios.[17] These benchmarking efforts underscore the complexities of clinical caption generation, including modality variation, clinical specificity, and vocabulary grounding.

The University of Murcia team achieved the top performance in the 2025 ImageCLEFmed captioning task.[18] Their approach leveraged the BLIP architecture,[19] which integrates a Vision Transformer (ViT) encoder with a language model decoder for medical image synthesis. The researchers fine-tuned a general-purpose pre-trained model, optimizing its performance through relevance-based selection. This methodology achieved an ROUGE-1 score of 0.259, establishing a current state-of-the-art benchmark, although direct comparisons are limited by variations across test sets.

Despite the progress enabled by foundation models and multimodal pretraining, challenges persist in generating clinically accurate and contextually rich captions. Many existing evaluations rely primarily on automated natural language generation metrics, which may not fully capture diagnostic relevance or narrative coherence. Moreover, the impact of prompt design, zero-shot, and few-shot conditions on model reliability has not been comprehensively explored across modern generative models in a single benchmark. This gap motivates the present study, which systematically evaluates state-of-the-art generative AI models—including GPT-4 and Google's Gemini 2.5—on the large-scale ROCOv2 dataset under diverse prompting strategies to better understand their capabilities and limitations in medical image captioning tasks.

3. METHODS

To evaluate generative AI models for medical image captioning, a benchmarking framework was designed, inspired by the ImageCLEFmed Caption task,[8] which has historically focused on generating descriptive captions for biomedical images using curated datasets and standardised evaluation metrics. This study extends that paradigm by applying it to foundation models and a radiology-specific dataset, ROCOv2.[9]

3.1 Dataset

The ROCOv2 dataset[9] is used as the primary evaluation resource. It is a large-scale multimodal dataset containing over 80,000 radiology images paired with expert-generated captions. ROCOv2 builds upon the original ROCO dataset used in ImageCLEFmed Caption, offering improved diversity, quality, and clinical relevance in its annotations. The experimental design choices were informed by preliminary evaluations conducted on subsets of 300 and 1,000 images sampled from the validation split. Final performance metrics reported in Tables 1 and 3 were computed on the ROCOv2 test set, which comprises 9,927 captioned images.

3.2 Models Evaluated

To assess the capabilities of current generative AI models for radiology image captioning, we selected a set of large-scale foundation models known for their natural language generation. All models were evaluated on ROCOv2 without additional fine-tuning, using zero-shot or few-shot configurations. The models that will be evaluated in this study include:

- **OpenAI GPT-4o:** Accessed via the OpenAI API, GPT-4o represents one of the most advanced publicly available large language models. It was evaluated under multiple prompting strategies, including domain-specific and retrieval-augmented prompts. GPT-4o has been widely used in general natural language understanding and generation tasks, but has not been fine-tuned specifically for biomedical imaging tasks.
- **Anthropic Claude Sonnet 4.5:** Claude Sonnet 4.5 is a state-of-the-art generative model developed by Anthropic. We evaluated Claude Sonnet 4.5 under similar prompting conditions as GPT-4o to provide a comparative analysis across model providers.

- **Google Gemini 2.5 Pro:** Gemini 2.5 Pro is a large-scale multimodal foundation model developed by Google. Preliminary experiments on a smaller validation subset showed stronger performance than other Gemini variants, motivating its selection for the full evaluation. As with the other models, Gemini 2.5 Pro was evaluated without task-specific fine-tuning.
- **Meta LLaMA 4:** two variants of Meta's LLaMA 4 architecture were employed: Llama 4 Maverick and Llama 4 Scout, both of which are open-weight models with publicly available parameters. While LLaMA models are not domain-specific, their open-weight nature enables controlled evaluation and potential fine-tuning in future work.

Each model is evaluated under the same prompting condition, with structured instructions, using a domain-specific prompt. In the RAG (retrieval-augmented generation) setting, examples from similar radiological cases were provided as contextual references during inference. This approach aimed to simulate the benefits of clinical knowledge retrieval in real-world settings and assess the model's ability to integrate multimodal context. By comparing these models across consistent experimental settings, we aim to highlight both strengths and limitations in their current ability to perform radiology-focused captioning, identify clinically relevant findings, and adhere to structured medical terminology.

3.3 Evaluation Metrics

Model performance is evaluated using a suite of automated metrics: BLEU (n-gram precision),[20] ROUGE (recall-oriented overlap),[21] METEOR (alignment-based penalty),[22] CIDEr (TF-IDF weighted consensus),[23] and BERTScore (semantic similarity via contextual embeddings).[24]

In medical image captioning, no single automatic metric is sufficient to fully characterize caption quality. Lexical overlap metrics capture surface-form fidelity and clinical terminology, while more flexible similarity measures account for paraphrasing and semantic equivalence. Therefore, a multi-metric evaluation strategy is adopted to jointly assess lexical accuracy, content coverage, and semantic consistency, ensuring both interpretability and comparability with established benchmarks.

By maintaining ROUGE-1 as a core metric, we ensure a direct point of comparison with the 10-year benchmarking legacy of ImageCLEF. The additional metrics, specifically CIDEr and BERTScore, provide a more granular assessment of how well models like Gemini and LLaMA 4 bridge the gap between surface-level lexical matching and clinical reasoning. Evaluations are conducted on the ROCOv2 test split using standardized preprocessing and publicly available implementations (e.g., the MS COCO toolkit and Hugging Face `evaluate`).

3.4 Prompt Engineering

Prompt design plays a critical role in adapting general-purpose generative models to radiology tasks. In this study, we compare multiple prompting strategies—baseline, which is a domain-specific prompt, and few-shot using RAG—to evaluate their impact on report caption quality.

Domain-Specific Prompt: Simulates radiologist expertise by emphasizing anatomical structure, clinical findings, and structured output. For example:

Listing 1: Domain-specific system prompt example for radiological image interpretation.

```
You are a radiology specialist reviewing a medical image for educational purposes.
    Your job is to provide a detailed and accurate caption for the medical image.

Use these instructions to guide your response:
- Your caption should be assertive, concise, and clinically accurate
- Focus on describing the visible anatomical structures, clinical findings, and
    medical devices
- Include any abnormalities, diagnostic findings, or interventions visible in the
    image
- Use appropriate medical terminology
- Be specific about the imaging modality, body region, and orientation when applicable
```

```
Following these steps:
Step 1: Analyze the image carefully and identify all visible anatomical structures,
    clinical findings, and medical devices
Step 2: Note any abnormalities, diagnostic findings, or interventions visible in the
    image
Step 3: Generate a clear and informative caption that describes the image
    comprehensively

Provide your response as a JSON object with the following structure:

{
    "caption": "A detailed and clinically accurate description of the medical image"
}
```

Retrieval-Augmented Prompting (RAG): To enhance contextual grounding, we incorporate examples from similar cases into the prompt leveraging a vector database to retrieve the n most similar images from the training set using RoBERTa[25] embeddings for similarity computation; in this study we use $n = 3$, since in preliminary tests this number proved to be a good balance between the cost of API usage and the increase in performance. These retrieved samples are then formatted as in-context examples to guide the language model during generation.

Listing 2: Retrieval-Augmented Generation (RAG) prompt structure example.

```
... <previous domain specific prompt>
Here are some similar medical images and their captions as examples to guide your
    response style and format
Example 1: {
  "image": <encoded image>
  "caption": "CT scan showing diffuse interstitial infiltrates, suggestive of
     pulmonary fibrosis."
}
... <other examples>
```

The model is then instructed to respond in the same format, fostering stylistic and semantic alignment with prior cases.

4. RESULTS

We compare two prompting strategies: (i) **Baseline**, a domain-specific prompt with detailed clinical framing and structured output with no examples to guide the model; and (ii) **RAG**, which augments the baseline prompt with retrieved image-caption examples to guide response style and content.

As shown in Tables 1 and 3, retrieval-augmented generation (RAG) consistently improves performance across all evaluated models. In particular, a controlled comparison between RAG and non-RAG configurations was conducted for the three best-performing RAG models, namely Gemini 2.5 Pro, LLaMA 4 Maverick, and GPT-4o.

For Gemini 2.5 Pro, the introduction of retrieval yields systematic gains across all evaluation metrics. BLEU-4 increases from 0.025 to 0.035, while ROUGE-L improves from 0.149 to 0.186, indicating enhanced sentence-level structure and improved alignment with ground-truth captions. Notably, the gains are more pronounced for higher-order n-gram metrics (BLEU-3 and BLEU-4), suggesting that retrieval primarily enhances compositional coherence and phrase-level structure, rather than isolated lexical overlap. This pattern indicates that retrieved examples serve as structural priors, guiding the model toward more fluent, well-formed captions.

Consistent improvements across recall and alignment-oriented metrics, including ROUGE-1, ROUGE-2, ROUGE-L, and METEOR, further indicate that RAG improves semantic coverage and sequence-level correspondence with the ground-truth annotations. In contrast, the more modest relative gains in METEOR compared to

Model \ Metric	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CIDEr
Gemini 2.5 Pro	0.155	0.078	0.043	0.025	0.200	0.057	0.149	0.242	0.012
Gemini 2.5 Pro (RAG)	0.203	0.104	0.059	0.035	0.241	**0.074**	0.186	0.257	0.063
LLaMA 4 Maverick	0.167	0.080	0.042	0.025	0.209	0.061	0.162	0.225	0.040
LLaMA 4 Maverick (RAG)	**0.218**	**0.107**	**0.060**	**0.037**	**0.250**	**0.074**	**0.198**	0.233	0.124
LLaMA 4 Scout (RAG)	0.191	0.091	0.049	0.029	0.220	0.060	0.171	0.226	0.061
GPT 4o	0.126	0.055	0.025	0.014	0.169	0.043	0.129	0.199	0.112
GPT 4o (RAG)	0.208	0.098	0.053	0.032	0.240	0.068	0.189	**0.275**	**0.134**
Claude 4.5 Sonnet (RAG)	0.183	0.085	0.045	0.027	0.211	0.055	0.164	0.211	0.065
Best ImageCLEFmed 2025	—	—	—	—	**0.259**	—	—	—	—

Table 1: Performance of multimodal foundation models on the ROCOv2 test set, comparing zero-shot and retrieval-augmented generation (RAG) configurations across automated metrics. The ImageCLEFmed 2025 benchmark is included for context, representing the state-of-the-art performance reported on an unseen test set.

Category	Metrics
Precision-Oriented Lexical Matching	BLEU-1...4
Recall-Oriented Lexical Coverage	ROUGE-1...2
Structural Sequence Overlap	ROUGE-L
Alignment / Coverage	METEOR
Consensus & Specificity	CIDEr
Semantic Similarity	BERTScore (Precision, Recall, F1)

Table 2: Categorization of text generation evaluation metrics based on their primary evaluation focus.

Model	BERTScore Precision	BERTScore Recall	BERTScore F1
Gemini 2.5 Pro	0.507	0.625	0.557
Gemini 2.5 Pro (RAG)	0.543	**0.628**	0.580
LLaMA 4 Maverick	0.521	0.595	0.552
LLaMA 4 Maverick (RAG)	**0.573**	0.599	**0.583**
LLaMA 4 Scout (RAG)	0.499	0.581	0.535
GPT 4o	0.488	0.568	0.523
GPT 4o (RAG)	0.563	0.598	0.578
Claude 4.5 Sonnet (RAG)	0.546	0.595	0.567

Table 3: Semantic similarity evaluation on the ROCOv2 test set.

ROUGE suggest that stylistic variation and paraphrasing remain challenging, potentially due to domain-specific language or annotation conventions.

The most substantial relative improvement for Gemini 2.5 Pro is observed in CIDEr, which increases from 0.012 to 0.063, representing more than a fivefold improvement. Since CIDEr emphasizes consensus among references while penalizing generic or boilerplate descriptions, this result suggests that retrieval effectively mitigates generic captioning behavior and encourages outputs that better match the dataset's reference distribution.

A qualitative inspection of the generated captions further reveals that, without retrieval, the model occasionally fails to match the desired tone or produces overly verbose descriptions. In particular, Gemini-generated captions tend to be longer than the ground-truth captions, which may negatively affect precision-oriented metrics. The inclusion of retrieved examples appears to alleviate these issues by constraining generation length and improving contextual alignment.

Similar trends are observed for the LLaMA 4 models, reinforcing the general effectiveness of retrieval-augmented generation in medical image captioning. For LLaMA 4 Maverick, the use of RAG increases CIDEr from 0.040 to 0.124, alongside improvements in BLEU-4 (from 0.025 to 0.037) and ROUGE-L (from 0.162 to 0.198). These gains indicate a clear shift away from generic or overly verbose descriptions toward captions that better reflect clinically relevant phrasing and reference structure. As with Gemini, the strongest improvements are observed in higher-order metrics, suggesting that retrieval provides semantic and structural guidance that supports more coherent and informative caption generation.

OpenAI's ChatGPT-4o delivered strong performance, achieving a ROUGE-L score of 0.189 and the highest CIDEr score of 0.134. Notably, it was among the models that benefited most from using a RAG strategy, recording a net ROUGE-L improvement of 0.06 and substantial gains across all three BERTScore metrics.

The Claude 4.5 Sonnet (RAG) configuration achieved competitive but relatively moderate performance on the benchmark, with ROUGE-L = 0.164 and CIDEr = 0.065. While these scores exceed those of some non-retrieval baselines, they remain below those of other models with RAG, particularly on metrics that emphasize semantic consensus and content relevance.

Despite these improvements, a more systematic analysis of error cases is necessary to better characterize the remaining failure modes. Such an analysis could inform targeted prompt refinements or additional control mechanisms to improve tone consistency, level of detail, and stylistic alignment, thereby further enhancing the robustness of retrieval-augmented caption generation.

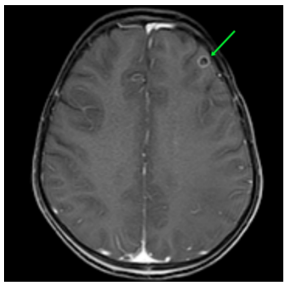
Table 3 reports BERTScore precision, recall, and F1 for all evaluated models. Overall, RAG consistently improves BERTScore F1 across architectures, indicating stronger semantic alignment between generated captions and reference descriptions. These gains are primarily driven by increases in precision, suggesting that RAG reduces the introduction of semantically irrelevant or spurious content while preserving recall.

Prediction Example: ROCOv2_2023_test_002186

Captions Comparison

Ground Truth Caption

Axial T1-weighted post-contrast image Incidental finding of a small peripherally enhancing neurocysticercosis lesion in left frontal lobe (green arrow).



Predicted Captions (all models)

Model	Caption
Gemini 2.5 Pro (RAG, n=3)	Axial T1-weighted post-contrast brain MRI demonstrating a ring-enhancing lesion in the left frontal lobe (green arrow), consistent with a brain abscess or metastasis.
Gemini 2.5 Pro (w/o RAG)	Axial T1-weighted post-contrast MRI of the brain demonstrating a ring-enhancing lesion in the left frontal lobe (green arrow), consistent with a cerebral abscess or metastatic deposit. The lesion exhibits a well-defined, thin, enhancing rim with a central area of low signal intensity, suggestive of necrosis or pus. There is minimal surrounding vasogenic edema. Additional smaller enhancing nodules are noted along the leptomeninges, particularly in the left cerebral hemisphere.
GPT-4o (RAG, n=3)	Axial T1-weighted brain MRI with contrast showing a small, round, enhancing lesion in the right frontal lobe cortex (green arrow), suggestive of a possible meningioma or metastatic lesion.
GPT-4o (w/o RAG)	This MRI scan of the brain shows a transverse section with a small, round lesion located in the right frontal lobe, indicated by a green arrow. The lesion appears hyperintense, suggesting the presence of a potential cyst or other fluid-filled structure. The surrounding brain tissue appears normal, with no significant edema or mass effect visible in this image.
LLaMA 4 Maverick (RAG, n=3)	Axial view of a contrast-enhanced MRI brain scan showing a ring-enhancing lesion in the left frontal lobe (green arrow).
LLaMA 4 Maverick (w/o RAG)	An axial T1-weighted MRI scan with contrast enhancement showing a ring-enhancing lesion in the left frontal lobe, indicated by the green arrow. The lesion has a characteristic 'ring' appearance due to peripheral enhancement with central necrosis, suggestive of an abscess or a tumor. The surrounding brain tissue appears to be mildly edematous. The image is likely from a patient with a suspected intracranial infection or neoplasm.
LLaMA 4 Scout (RAG, n=3)	Axial view brain MRI showing a small ring-enhancing lesion in the left frontal lobe (green arrow).
Claude Sonnet 4.5 (RAG, n=3)	Axial T1-weighted brain MRI demonstrating a small round hyperintense lesion in the right frontal cortex (green arrow), consistent with a calcified granuloma or cavernoma

Figure 1: Comparison of ground-truth and predicted captions for an illustrative ROCOv2 test image across multiple foundation models evaluated under both zero-shot and retrieval-augmented generation (RAG) conditions.

Among the evaluated models, LLaMA 4 Maverick (RAG) achieves the highest BERTScore F1 (0.583), followed closely by Gemini 2.5 Pro (RAG) (0.580) and GPT-4o (RAG) (0.578), reflecting a consistent benefit of retrieval augmentation across both open and proprietary models. In contrast, non-RAG configurations exhibit lower F1 scores, particularly for GPT-4o and LLaMA 4 Maverick, reinforcing the role of retrieval in improving semantic fidelity.

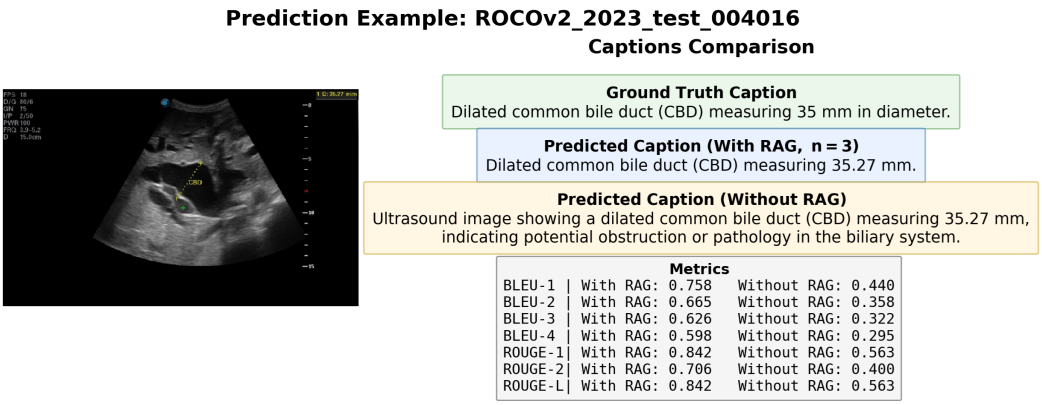


Figure 2: Qualitative captioning results for a ROCOv2 test sample using the LLaMA 4 Maverick model, including the input image, ground truth caption, predictions with and without RAG, and the corresponding BLEU and ROUGE metrics.

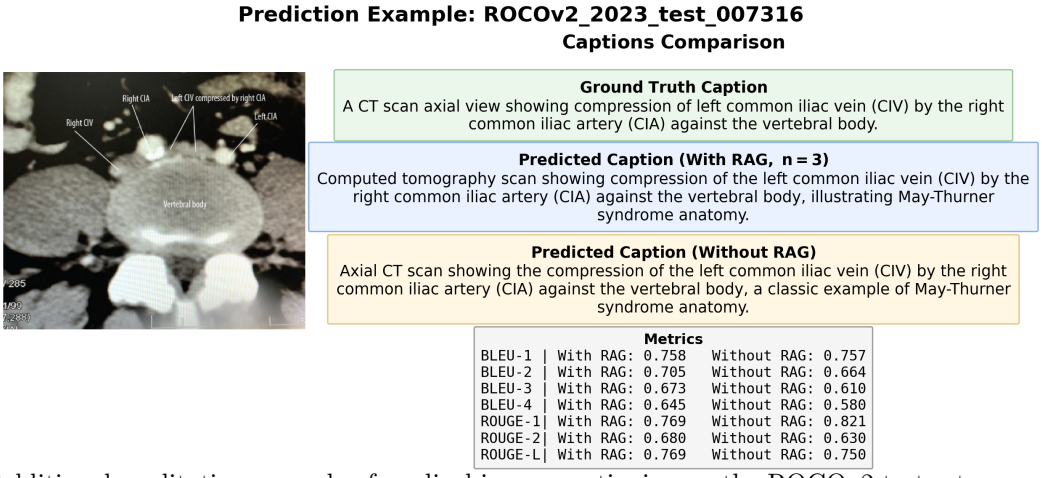


Figure 3: Additional qualitative example of medical image captioning on the ROCOv2 test set generated by the LLaMA 4 Maverick model, showing ground truth annotations, predicted captions with and without RAG, and associated BLEU and ROUGE scores.

A common pattern across models is that recall values are systematically higher than precision, indicating that most clinically relevant content present in the references is generally covered by the generated captions, while precision remains more sensitive to extraneous or loosely related semantic content. The observed precision improvements under RAG therefore suggest better control over content specificity without sacrificing coverage.

Figure 1 provides a qualitative overview of the captioning behavior across all evaluated models for a single ROCOv2 test image, presenting the ground-truth caption alongside predictions generated with and without retrieval-augmented generation (RAG). The comparison highlights substantial variability in both descriptive focus and clinical precision among models. While most approaches correctly identify a ring-enhancing lesion in the frontal lobe, notable differences emerge in anatomical localization, diagnostic interpretation, and verbosity. RAG-enabled models generally produce more concise and radiologically grounded descriptions, whereas non-RAG variants tend to generate longer captions, occasionally introducing speculative findings or clinically unsupported details. This figure illustrates how retrieval influences not only stylistic aspects but also diagnostic interpretation.

To provide a deeper understanding of the performance gains observed in Table 1, Figures 2 and 3 present illustrative qualitative results from the LLaMA 4 Maverick model under non-RAG and RAG settings. This model was selected for detailed visualization as it represents the highest-performing open-weight architecture in our study, nearing the current state-of-the-art ROUGE-1 benchmark.

A comparative analysis of these figures demonstrates that the RAG-enhanced configuration consistently yields

more concise clinically focused descriptions. This configuration better aligns with the professional radiological reporting style by leveraging the structural and semantic priors from the retrieved examples. In contrast, the non-RAG variant often produces verbose, redundant outputs that include speculative findings.

This qualitative pattern is also reflected in the example-level evaluation metrics shown in the figures. For the ROCOv2 test case `004016`, the RAG-based prediction achieves substantially higher BLEU scores across all n-gram orders as well as markedly improved ROUGE values, indicating stronger alignment with the reference caption at both lexical and semantic levels. For test case `007316`, although the RAG configuration achieves slightly higher scores in some higher-order metrics, the absolute differences are minimal (e.g., BLEU-1 of 0.758 with RAG versus 0.757 without RAG, and ROUGE-L of 0.769 versus 0.750). Consistent with these results, the captions generated by both approaches are semantically and structurally similar, differing only in minor phrasing choices. These findings indicate that, for this specific example, the benefits of RAG are marginal and are primarily reflected in structural and phrase-level metrics rather than in substantial improvements in surface-level lexical matching.

5. DISCUSSION

The results of this benchmark study provide a critical assessment of the current state of MFMs in the specialized domain of radiology. As the field marks the 10th anniversary of the ImageCLEFmed Caption task, our findings illustrate a significant shift from the concept-detection architectures of the previous decade toward systems capable of nuanced, narrative-driven clinical reasoning. A primary finding of this research is that RAG consistently enhances performance across all linguistic and clinical metrics. The most remarkable gain was observed in the CIDEr metric, which increased by more than a factor of five for the Gemini model when RAG was enabled, suggesting that retrieval helps mitigate generic captioning behavior and encourages outputs that better match the dataset's reference distribution.

Our analysis suggests that retrieved examples act as structural priors, guiding models like Gemini 2.5 Pro, GPT-4o and LLAMA 4 Maverick to move beyond isolated lexical overlap toward more coherent, phrase-level structures. Notably, the LLAMA 4 Maverick (RAG) configuration demonstrated exceptional utility, achieving a ROUGE-1 score of 0.250. This performance nearly matches the 0.259 benchmark established by the UMUTeam at ImageCLEF 2025, which used specialized, fine-tuned BLIP architectures. The performance gap between variants is also notable; for instance, LLaMA 4 Maverick (RAG) achieved a CIDEr score of 0.124, nearly doubling the 0.061 achieved by the LLaMA 4 Scout (RAG) variant, underscoring the influence of model scale even within the same architecture family. Furthermore, the BERTScore analysis reveals that semantic gains are primarily driven by increased precision, suggesting that RAG effectively reduces the introduction of semantically irrelevant or spurious content—a critical factor for clinical safety. The fact that an open-source model like LLAMA 4 can achieve state-of-the-art results through advanced prompting alone, without task-specific fine-tuning or proprietary API costs, has significant implications for the clinical community. Such models offer a cost-effective and transparent alternative to proprietary systems, potentially lowering the barrier to integrating multimodal AI into real-world hospital infrastructures, where data privacy and resource allocation are paramount.

However, several challenges persist that limit the immediate integration of these models into clinical workflows. While RAG improves semantic coverage, the more modest relative gains in METEOR compared to ROUGE suggest that capturing the specific stylistic nuances and varied medical vocabulary of human radiologists remains difficult for MFMs. Furthermore, Gemini-generated captions tended to be longer than the ground-truth references, which can negatively affect precision-oriented metrics and clinical utility. Despite these limitations, the transition to the ROCOv2 dataset—an expert-annotated successor to the original ROCO—provides the necessary complexity to test whether foundation models can finally bridge the gap between automated captioning and professional radiological standards. Ultimately, this work demonstrates that the focus must shift from mere concept detection to interpretable multimodal reasoning supported by domain-aware prompt engineering.

Future work should investigate prompting strategies that reduce response verbosity without additional examples, establishing guardrails for output length while prioritizing descriptive analysis and restricting diagnostic inferences to high-certainty cases. Additionally, the impact of fine-tuning using the available training dataset

should be evaluated and compared against Retrieval-Augmented Generation (RAG) performance. Further research will also explore ensemble methods combining multiple foundation models (MFMs) and agentic frameworks to simulate collaborative differential diagnosis workflows among large language models.

6. CONCLUSIONS

This study presents a comprehensive evaluation of state-of-the-art generative AI models for medical image captioning, with a focus on the radiology domain. By benchmarking large-scale foundation models on the ROCOv2 dataset and following evaluation protocols established in the ImageCLEFmed Caption challenge within the CLEF framework, this work highlights both the transformative potential and the current limitations of generative AI in generating clinically accurate and contextually relevant captions. Our analysis reveals that RAG is a vital catalyst for performance, consistently enhancing clinical specificity and improving metrics such as CIDEr, which increased by over fivefold in our evaluations.

Qualitative findings from Figures 1,2 and 3 corroborate these metrics, illustrating that RAG produces more concise and radiologically grounded descriptions compared to the often verbose and speculative zero-shot outputs. The results also demonstrate strong semantic alignment across providers, with BERTScore F1 values up to 0.583, confirming that generated captions preserve medical meaning beyond surface-level lexical matching . Notably, open-weight models like LLAMA 4 Maverick, when using RAG, achieved a ROUGE-1 score of 0.250, approaching the 2025 state-of-the-art benchmark of 0.259 established by specialized architectures.

However, preliminary observations suggest that clinical accuracy and consistency vary with image complexity and prompt specificity. Domain-specific prompt engineering appears to enhance output quality, underscoring the importance of tailored interaction strategies in medical applications. While RAG improves semantic coverage, challenges regarding model verbosity, particularly with Gemini, and the accurate capture of stylistic nuances inherent in expert radiological reporting persist. Future research will involve expert clinical reviews to assess the diagnostic correctness of these models and a targeted analysis of failure modes across diverse anatomical complexities. Ultimately, this work contributes to the development of more reliable and explainable multimodal AI, offering a clear perspective on the evolving role of generative models as medical imaging enters its second decade of standardized evaluation.

ACKNOWLEDGMENTS

This work was partly supported by the projects GRESEL-UNED (PID2023-151280OB-C22) funded by MICIU/AEI/ AEI 501100011033 and ANNOTATE (PID2024-156022OB-C31) funded by MICIU/AEI/10.13039/ 501100011033 and the European Social Fund Plus (ESF+).

REFERENCES

[1] Xiao, X., Zhang, Y., Nguyen, T.-H., Lam, B.-T., Wang, J., Zhao, L., Hamm, J., Wang, T., Li, X., Wang, X., Xu, H., Liu, T., and Xu, M., "Describe anything in medical images," (2025).

[2] Lee, H., Cho, H., Park, J., Chae, J., and Kim, J., "Cross encoder-decoder transformer with global-local visual extractor for medical image captioning," *Sensors* **22**(4) (2022).

[3] Reale-Nosei, G., Amador-Domínguez, E., and Serrano, E., "From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation," *Medical Image Analysis* **97**, 103264 (2024).

[4] Lee, J.-O., Zhou, H.-Y., Berzin, T. M., Sodickson, D. K., and Rajpurkar, P., "Multimodal generative ai for interpreting 3d medical images and videos," *npj Digital Medicine* **8**, 273 (may 2025).

[5] Rao, V. M., Hla, M., Moor, M., Adithan, S., Kwak, S., Topol, E. J., and Rajpurkar, P., "Multimodal generative ai for medical image interpretation," *Nature* **639**, 888–896 (mar 2025).

[6] García Seco de Herrera, A., Yagis, E., Pinpo, N., Abolghasemi, V., Andritsch, J., Chaichulee, S., Dicente Cid, Y., and Ingviya, T., "Ensemble deep learning architectures for detecting pulmonary tuberculosis in chest x-rays," *Scientific Reports* **16**(1), 1242 (2026).

[7] Ferro, N. and Peters, C., eds., [*Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*], vol. 41 of *The Information Retrieval Series*, Springer Cham (2019).

[8] Damm, H., Pakull, T. M. G., Becker, H., Bracke, B., Eryilmaz, B., Bloch, L., Brüngel, R., Schmidt, C. S., Rückert, J., Pelka, O., Schäfer, H., Idrissi-Yaghir, A., Ben Abacha, A., Garc'Seco de Herrera, A., Müller, H., and Friedrich, C. M., "Overview of ImageCLEFmedical 2025 – medical concept detection and interpretable caption generation," in [*CLEF 2025 Working Notes*], *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain (September 9–12 2025).

[9] Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C. S., Koitka, S., Pelka, O., Ben Abacha, A., García Seco de Herrera, A., Müller, H., Horn, P. A., Nensa, F., and Friedrich, C. M., "ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset," *Scientific Data* **11**(1), 688 (2024).

[10] Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., and Li, S.-N., "Medical image captioning via generative pretrained transformers," *Scientific Reports* **13**(1), 4171 (2023).

[11] Moon, J. H., Lee, H., Shin, W., Kim, Y.-H., and Choi, E., "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE Journal of Biomedical and Health Informatics* **26**(12), 6070–6080 (2022).

[12] Li, Z., Chen, J., Xu, Y., Wang, Z., Li, P., Zhang, J., Zhang, W., Chen, Y., Liu, M., Wu, S., et al., "Medblip: Bootstrapping language-image pre-training from 3d medical images and texts," *arXiv preprint arXiv:2305.10799* (2023).

[13] Zhao, Y., Braytee, A., and Prasad, M., "Dualprompt-medcap: A dual-prompt enhanced approach for medical image captioning," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*], –.

[14] Tran, T. T., Nguyen, D. V. M., and Huynh, H. T., "A novel transformer-based framework for chest x-ray captioning with clinical sign detection," *SN Comput. Sci.* **6**, 520 (2025).

[15] Cherukuri, T. K., Shaik, N. S., Bodapati, J. D., and Ye, D. H., "Gcs-m3vlt: Guided context self-attention based multi-modal medical vision language transformer for retinal image captioning," in [*ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 1–5 (2025).

[16] Kamal, M. S., Nimmy, S. F., Islam, M. R., and Naseem, U., "Explainable medical image captioning," in [*Companion Proceedings of the ACM on Web Conference 2025*], *WWW '25*, 2253–2261, Association for Computing Machinery, New York, NY, USA (2025).

[17] Hoque, M., Chowdhury, R. N., Hasan, M. R., Peter, O. O. E., Khalifa, F., and Rahman, M. M., "An empirical evaluation of low-rank adapted vision–language models for radiology image captioning," *Bioengineering* **12**(12) (2025).

[18] Pan, R., Bernal Beltrán, T., García Díaz, J. A., and Valencia-García, R., "UMUTeam at ImageCLEF 2025: Fine-tuning a vision-language model for medical image captioning and SapBERT-based reranking for concept detection," in [*CLEF2025 Working Notes*], *CEUR Workshop Proceedings*, CEUR-WS.org, Madrid, Spain (September 9-12 2025).

[19] Li, J., Li, D., Xiong, C., and Hoi, S., "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in [*International Conference on Machine Learning*], 12888–12900 (2022).

[20] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., "Bleu: a method for automatic evaluation of machine translation," in [*Proceedings of the 40th annual meeting of the Association for Computational Linguistics*], 311–318 (2002).

[21] Lin, C.-Y., "Rouge: A package for automatic evaluation of summaries," in [*Text summarization branches out*], 74–81 (2004).

[22] Lavie, A. and Denkowski, M. J., "The meteor metric for automatic evaluation of machine translation," *Machine translation* **23**(2), 105–115 (2009).

[23] Vedantam, R., Lawrence Zitnick, C., and Parikh, D., "Cider: Consensus-based image description evaluation," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 4566–4575 (2015).

[24] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y., "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675* (2019).

[25] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V., "Roberta: A robustly optimized bert pretraining approach," (2019).