AI-driven Mammogram Report Generation with Open-source VLMs and RAG Prompting

Raiyan Jahangir, Nafiz Imtiaz Khan, Anubhav Mishra

Abstract

Breast cancer is a leading cause of mortality among women, which accentuates the importance of early and accurate diagnosis. Automating mammogram report generation with AI-powered Vision Language Models (VLMs) can enhance radiologist efficiency, reduce workload, and improve diagnostic consistency. In this project, we explored the use of open-source Vision Language Models (VLMs) for automated mammogram report generation. We experimented with different techniques/configurations: iterative prompt tuning, retrieval-augmented generation (RAG), and model fine-tuning. We evaluate three models—Llava, Mistral, and Qwen2.5—across different prompting methods, finding that Qwen2.5 performs best in zero/few-shot settings, while Llava excels when integrated with RAG. Our findings suggest that local VLMs, particularly with RAG-based enhancement, can effectively assist radiologists in generating structured mammogram reports.

1 Introduction & Background

One of the leading causes of death in women globally is breast cancer [1]. It constitutes 30% of all the new cancer cases in women every year [2]. To detect any cancer at an early stage, it is necessary for women to screen and test for breast cancer at regular intervals. Mammography is a modern technique that is used for detecting breast cancers. It uses X-rays to capture images of breast tissue, widely known as mammogram images [3]. The radiologists inspect the mammogram images and look for any potential abnormalities that might lead to cancer. Based on their observations, they write a report on the current condition of the breast tissue and its compositions, any abnormal findings and a Breast Imaging Reporting and Data System (BIRADS) score, a type of severity score which range from 0 to 6 [4]. Based on this BIRADS score, the radiologists also mention necessary treatments and further actions in the mammogram report. To ease the work of radiologists, researchers have been trying to incorporate AI's benefits into healthcare. With the development of Large Language Models (LLMs) [5], and most recently, the Vision Language Models (VLMs) [6], the importance of AI assistance in healthcare has increased manifold. VLMs with their multimodal ability, can read both texts and images together and generate human-like text responses based on the input. The application of VLM in mammogram report generation can help ease the process of determining abnormalities in breast tissues. This will help and assist the radiologists in their inspection and help them decide on their results and findings much quicker.

Making VLMs carry out specialized tasks like generating reports from mammogram images involves tuning the models [7]. The models are pre-trained to answer generalized questions. To make them carry out specialized tasks, they have to be either provided with tailored prompts to answer in a specific manner [8] or pre-trained with a specific dataset, which is known as fine-tuning [9]. Each tuning method has pros and cons and trade-offs between complexity and performance. Prompt tuning might be suitable for cases where there are hardware limitations. It helps the models to generate responses in a proper format. However, the model weights never change in prompt tuning and so the models actually do not learn anything. On the other hand, fine-tuning requires a lot of data and intensive computational and hardware resources. The popular and powerful models like ChatGPT mostly comes with high monetary costs and so they are not accessible to all people. Again, they are trained to answer generalized question and has a tendency to hallucinate answers [10]. In such cases, open-source VLMs can be a possible solution. They can be run locally and accessible without any financial transactions.

A Retrieval Augmented Generation (RAG) framework can be used to enhance a VLM's response by providing relevant information to it as a context along with user queries [11]. This can be helpful to a model so that the model always generates its answers from a relevant source instead of answering on its

own. RAG with in-context learning [12] can be an effective way to generate mammogram reports which will be without any hallucinations. Very few research explored how VLMs can be tuned to generate reports of mammogram images and which method and model would be the most suitable in a RAG framework.

Research is already being conducted in this area to determine how VLMs can be made useful to aid radiologists in analyzing mammogram images. Ghosh et al. [13] introduced Mammo-CLIP, a VLM pre-trained on mammogram-report pairs to improve data efficiency and robustness. They also introduced the Mammo-Factor, a novel feature attribution method for spatial feature interpretation. Jain et al. [14] proposed a multimodal model that integrated mammographic images with patients' clinical histories using Vision Transformers (ViT) and RoBERTa-based textual embeddings. While these approaches improved mammographic interpretation, they focused more on feature alignment and classification rather than generating comprehensive, context-aware radiology reports. Moura et al. [15] explored several existing VLM models and their capability to classify breast tissue density and BIRADS categorization highlighting their strengths and limitations in clinical settings. Khan et al. [16] implemented a VLM framework for creating unique case sets of mammograms by automating case selection using image-text retrieval. Cao et al. [17] introduced MammoVLM, a new VLM developed with GLM-4-9B LLM and visual encoder to generate diagnostic assistance based on mammogram images. They fine-tuned the model with their own dataset and showed that their model capability is equivalent to a junior radiologist.

Based on the above background review, our research work focuses on determining the best model with the best prompting method for automated mammogram report generation and ultimately aiding radiologists in making more informed and reliable diagnoses.

The research questions of this work are as follows:

RQ$_1$: How effectively can local VLMs generate mammogram reports given a mammogram image in presence or absence of tailored prompts?
RQ$_2$: Can RAG framework and fine-tuning improve the performance of report generation?

2 Methodology

The overall methodology of the project is shown in figure 1. It is discussed in detail in the following subsections:

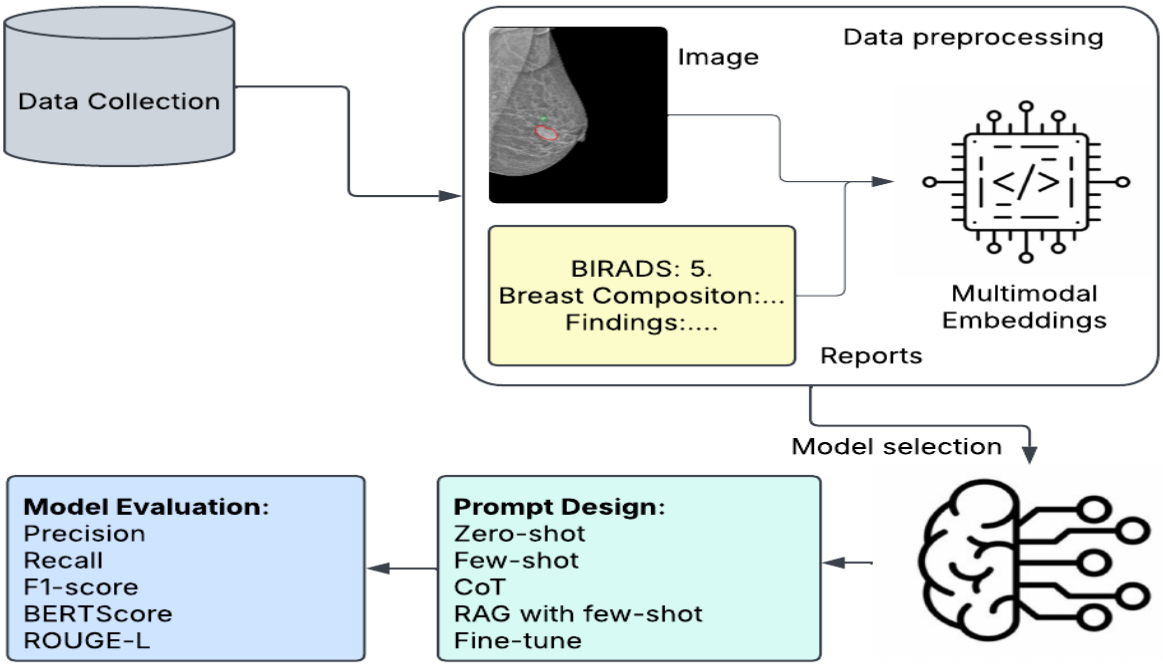

Figure 1: Overall methodology of the study

2.1 Data collection

We used the dataset prepared by Oza et al. [18] in our studies. The dataset consists of 510 mammogram images and 510 mammogram reports that corresponds to the mammogram images. Each report contains the breast composition, the BIRADS score and findings. There are also pixel-level annotations present for those images which have any forms of abnormalities. We use the pixel-level annotated images for the images with abnormalities and the normal images for healthy breasts.

2.2 Data synthesis

At first, we resize all the images to 224 x 224 to reduce memory overloading. After this, we convert each reports from text files to json files for easier input to models and easier outputs. In the process, we remove any unnecessary characters from the reports and clear any inconsistencies. For example, some reports had more than one BIRADS score provided. We kept the severe one in the processed report. We finally convert the clean reports to the json files.

2.3 Model selection

We selected the models which can take in images and texts as input and generate texts as outputs and can be run in GPUs with lower specifications. We selected Llava, Mistral-7B, and Qwen2.5 for our work. They are briefly described below:

2.3.1 Llava

Stands for Large Language And Vision Assistant. It was developed by Liu et al. [19] by combining a vision encoder with Vicuna, a fine-tuned LLama model [20] for general purpose visual and language understanding. The Llava model is shown to achieve 92.53% accuracy on Science QA dataset.

2.3.2 Mistral 7B

Mistral 7B is a 7-billion parameter model developed by Mistral AI [21]. This model is shown to have performed better in terms of both accuracy and efficiency in several benchmark datasets compared to other larger models.

2.3.3 Qwen2.5

Qwen2.5 [22] is an enhanced version of Qwen2 model. It also has 7 billion parameters. Although, initially developed to extract text and object from images, it has potential to identify abnormalities in mammogram images.

2.4 Prompt design

To enable an LLM to generate a proper response, it is necessary to feed it with an appropriate prompt [23]. The prompt provides the LLM with an additional context as well as to guide towards formatted output with clear, concise, relevant information. In our study, we focused on zero-shot [24], few-shot (n=2) [25] and chain of thought (CoT) [26] prompting technique. A small description of the propmting techniques used is described below:

2.4.1 Zero-shot

For zero-shot learning, we simply provided a small instruction on how the model will respond when provided with a question and how formatted will its answer be. The prompt is provided below:

> I will provide you with a mammogram image. Your task is to analyze the image and extract key diagnostic information, including breast composition, BIRADS category, and any significant findings. Present the output in a structured JSON format with the following keys: IMG_ID, Breast_Composition, BIRADS, and Findings. Ensure the response is precise, medically relevant, and well-organized. Please follow the below given JSON format for your response:
> {
> "IMG-ID": "<Image_Filename>",
> "BREAST-COMPOSITION": "<Description of breast tissue composition>",
> "BIRADS": "<BIRADS category; any values between 1 to 6. BI-RADS category is a standardized classification for breast imaging findings, ranging from 1 to 6, where: BI-RADS 1 indicates a negative result with no abnormalities; BI-RADS 2 signifies benign findings with no suspicion of cancer; BI-RADS 3 suggests a benign lesion, requiring short-term follow-up to confirm stability; BI-RADS 4 represents a suspicious abnormality needing biopsy, further divided into 4A (low suspicion), 4B (moderate suspicion), and 4C (high suspicion); BI-RADS 5 is highly suggestive of malignancy with a high probability of cancer; and BI-RADS 6 confirms a known malignancy with a biopsy-proven cancer diagnosis.>",
> "FINDINGS": "<Summary of any abnormalities, calcifications, or other observations>"
> }

2.4.2 Few-shot

For few-shot learning, we added two examples of answers that we expect the model to deliver along with the prompt that was provided in zero-shot learning. The modified prompt is shown below:

<Prompt from Zero-Shot> +

Here are some examples of doctor annoted reports to guide you:
Example 1:
{
“IMG-ID”: “IMG001.png”,
“BREAST-COMPOSITION”: “Predominantly fibro fatty breast parenchyma (ACR B)”,
“BIRADS”: “3”,
“FINDINGS”: “Small well defined soft nodular opacity- benign lesion (BIRADS-3). Benign and vascular calcifications. Skin and nipple - no abnormality. No axillary adenopathy”
}

Example 2:
{
“IMG-ID”: “IMG002.png”,
“BREAST-COMPOSITION”: “Fibro fatty with scattered glandular breast parenchyma (ACR B)”,
“BIRADS”: “1”,
“FINDINGS”: “No abnormal soft opacity. Vascular calcifications. Skin and nipple - no abnormality. Benign looking axillary adenopathy”
}

2.4.3 Chain-of-thought

For chain-of-thought prompting technique, we modify the context such that the model thinks itself as a radiologist. And based on how radiologists would normally inspect the mammogram images “thinking”, the model will also “think” and infer the results step by step. For example, a radiologist first checks the tissue density of the breasts. After deciding on the breast density, they look for any abnormal findings. Upon the inspection of this two, they conclude their inspection with the BIRADS score. We provide with a prompt so that they generate answers based on this format of thought. The chain-of-thought prompt template is shown below:

<Prompt from Zero-Shot> +

Step 1: First find out the Breast Density Category in ACR Format where
< Description of all sorts of Breast Density >

Step 2: Then determine any abnormal findings (or tumors) in the image. Findings are abnormalities or observations detected in a mammogram. Each type indicates different levels of concern:

< Description of all sorts of Findings >

Step 3: Now finally, determine the BIRADS of the mammogram where

< Description of BIRADS categories >

Step 4: Please follow the below given JSON format for your response:
{
"IMG-ID": "<Image_Filename>",
"BREAST-COMPOSITION": "<Description of breast tissue composition>",
"BIRADS": "<BIRADS category>",
"FINDINGS": "<Summary of any abnormalities, calcifications, or other observations>"
}

2.5 Experimental design

We carry all our experiments on a local machine platform. The temperature parameter for the models was always set to 0. Because, temperature is a hyperparameter that determines the creativeness of the model [27]. More the value of temperature, more the tendency of the model to generate random or creative answers. Since mammogram reports are sensitive and we only want the correct information, we set the temperature to the lowest value possible. The experimental setups are divided into following categories:

2.5.1 Base configuration

The base setting involves providing only the prompt to the model. We used the Ollama tool to acquire responses from the models locally. The responses from the models are saved in the local machine. The prompts were modified based on zero-shot, few-shot and chain-of-thought prompting techniques.

2.5.2 RAG configuration

We form a RAG pipeline to supply additional context to the LLM. The architecture of the RAG pipeline is provided in figure 2. It shows the components that are interconnected and how it effectively retrieves and leverages relevant information from a vector database.

For this, each of the 510 image-report pairs have been converted to embeddings. The image and text embeddings were carried out by the multimodal OpenCLIPEmbedding function that uses a Vision Transformer model [28] to extract the embeddings. The embeddings are then converted to semantic indexes and stored in ChromaDB, a vector database for similarity search. When a user queries, the query is also converted to an embedding with the embedding function and then sent to the database to extract 3 most similar indices. These 3 similar indices are then inserted into the prompt template. Thus, using

this procedure, we are able to retrieve a dynamic prompt template where the examples are always similar to the user query. This in turn, will help the model generate more accurate answers.

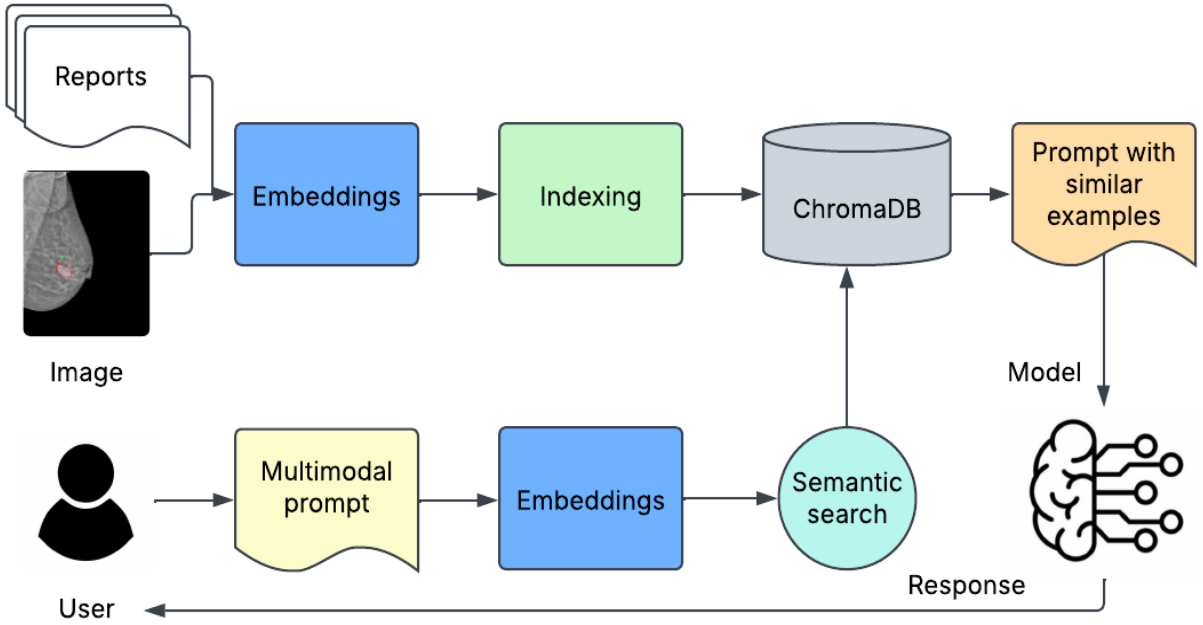

Figure 2: The RAG pipeline used in this study

2.5.3 Fine-tuning configuration

After running all the models and recording their performances, we fine-tune the best-performing model. For fine-tuning, the dataset is split into training, validation, and test set. Then, the model is trained. As the Qwen2.5 model was found to perform better, this particular model was fine-tuned. As training a VLM requires a powerful computer specification, we employ QLora [29], a parameter-efficient fine-tuning method to fine-tune the model on a training set (408 images) and validated on a validation set (51 images). The hyperparameters used in the model are given below:

Fine-tuning parameters:
Batch size: 1
Optimization: Qlora
Epochs: 50
Learning rate: 2e-4
Stopping criteria: Early Stopping

2.6 Evaluating model response

We evaluate the models based on the 3 types of outputs in the mammogram report namely the BIRADS, breast composition, and findings. We evaluate the performance on BIRADS based on precision, recall, and f1-score. The reason of selecting these parameters is because the BIRADS response will be either true or false like a classification task. For breast composition and findings, we apply BERTScore [30] and

ROUGE-L since these parameters focus on finding out the semantic meaning rather than simply classification task. BERTScore measures the semantic meaning among two sentences or paragraphs. ROUGE-L stands for Recall-Oriented Understudy for Gisting Evaluation. It measures the longest matching sequence of words using Longest Common Subsequence algorithm. Before recording the performance, we carry out 3-fold cross validation and average the results of each models in each prompting techniques.

3 Results

The performances of the models in all the prompting methods are shown in table 1. The 'X' mark on the fine-tuning section indicates that the experiment wasn't carried out. This section answers the research questions of this research.

RQ1: How effectively can local VLMs generate mammogram reports given a mammogram image in presence or absence of tailored prompts?

The base configuration was set for answering this RQ. It is noticeable from table 1 that Qwen2.5 model performs better than Mistral and Llava in zero-shot and few-shot techniques (above 0.5). Although, the ROUGE-L on findings and breast composition is relatively low (less than 0.4) for both prompting techniques. For chain-of-thought prompting technique, Mistral performed better than the other two. However, the problem os low ROUGE-L still persists. However, the reason for this could be because ROUGE-L focuses more on the order and sequence of words in a sentence rather than the semantic meaning itself. In our evaluation, we are more focused on the semantic meaning of output rather than the sequence of words chosen. Therefore, we can rightly said that local VLMs can be used effectively to generate mammogram reports with appropriate prompting techniques.

RQ_1 Findings: Local VLMs generate effective mammogram reports with tailored prompts, with Qwen2.5 excelling in zero/few-shot settings. Despite low ROUGE-L scores, semantic evaluation suggests their viability.

RQ2: Can RAG framework and fine-tuning improve the performance of report generation?

We have applied few-shot with RAG framework for all the models and finally fine-tuned the Qwen2.5 model. It is noticeable that the Llava model performs better than the other two when RAG has been applied to decide on the prompt examples. All of its performance paramaters are above 0.4. Even the ROUGE-L for findings and BIRADS are above 0.4, which can be considered a satsifactory performance. The Qwen2.5 model also has an impressive ROUGE-L (equal or above 0.4). However, the performance of Mistral remains mostly unchanged. The reason of better ROUGE-L could be deduced that with RAG, the model always looks for relevant and similar information from the vector database and so it catches the pattern of intended response easily.

The fine-tuning of Qwen2.5 however, does not show satisfactory performance. We assume that because of low specification of the local machine as well as the small dataset, it was difficult for the model to learn properly. However, with better specifications, it might perform better. Therefore, we can say that with the inclusion of RAG framework, the performance of mammogram report generation in models can be improved significantly.

RQ_2 Findings: The RAG framework improves mammogram report generation, with Llava and Qwen2.5 achieving satisfactory ROUGE-L scores. However, fine-tuning Qwen2.5 showed limited gains, likely due to limited training datset.

4 Limitations & Future Work

Our study has several limitations due to computational constraints. Firstly, we carried out out experiments in a relatively small dataset involving sample size of 510. Secondly, we could not try advanced

Prompt Method	Evaluation Parameters		Models		
			Llava	Mistral	Qwen2.5
Zero-Shot	BIRADS	Precision	0.53	0.52	0.59
		Recall	0.47	0.72	0.64
		F1-score	0.5	0.61	0.61
	Findings	BERTScore	0.54	0.56	0.58
		ROUGE-L	0.09	0.17	0.11
	Breast Composition	BERTScore	0.64	0.6	0.65
		ROUGE-L	0.1	0.1	0.14
Few-Shot	BIRADS	Precision	0.52	0.56	0.55
		Recall	0.72	0.28	0.68
		F1-score	0.61	0.13	0.6
	Findings	BERTScore	0.6	0.61	0.62
		ROUGE-L	0.29	0.19	0.32
	Breast Composition	BERTScore	0.86	0.87	0.88
		ROUGE-L	0.62	0.58	0.68
Chain-of-thought	BIRADS	Precision	0.55	0.66	0.65
		Recall	0.68	0.72	0.57
		F1-score	0.6	0.61	0.59
	Findings	BERTScore	0.39	0.55	0.55
		ROUGE-L	0.04	0.13	0.12
	Breast Composition	BERTScore	0.49	0.67	0.72
		ROUGE-L	0.07	0.16	0.2
RAG-Few-Shot	BIRADS	Precision	0.64	0.55	0.65
		Recall	0.69	0.62	0.61
		F1-score	0.65	0.58	0.63
	Findings	BERTScore	0.69	0.62	0.68
		ROUGE-L	0.49	0.26	0.4
	Breast Composition	BERTScore	0.83	0.84	0.88
		ROUGE-L	0.66	0.5	0.67
Fine-tuning	BIRADS	Precision			0.65
		Recall		X	0.35
		F1-score			0.3
	Findings	BERTScore		X	0.56
		ROUGE-L			0.14
	Breast Composition	BERTScore		X	0.36
		ROUGE-L			0

Table 1: Comprehensive overview of the performance of vision language models across five configurations: Zero shot, Few Shot, Chain-of-thought, RAG-with-Few-Shot, and Fine-tuning. In the Fine-tuning configuration, "X" indicates that a particular model was not fine-tuned.

prompting techniques like the Tree-of-Thought [31], Chain-of-thought [26], Reason+Action prompting [32]. Thirdly, we only tried 3 open-source VLMs for our studies. Fourthly, we couldn't try larger and more capable models and fine-tune them because of hardware limitations. Future works will focus on carrying out the research work on training larger models with larger datasets with advanced prompting techniques on better hardware platform.

5 Conclusion

This study showed the capability of open-source local VLMs in generating mammogram reports as substitutes to proprietary ones used for medical tasks. By addressing the research questions, it was shown that the Llava model in a RAG framework with few-shot prompting excelled in generating mammogram reports. Thus, these models can play an important role in helping the radiologists to come to an inspection decision and in writing and making mammogram reports.

6 Team Member Contributions

Raiyan: Data Collection, Data Pre-processing, Model Fine-tuning, Setting Up RAG Pipeline, Report Writing
Nafiz: Pipeline Development, Data Post-processing, Model Prompt tuning, Setting up RAG Pipeline, Report Writing
Anubhav: Report writing, Presentation Slide preparation

Code & Data Availability Statement

The project's codebase is available on GitHub at this link: https://github.com/Nafiz43/VLMs-for-Mammograms

References

[1] N. Azamjah, Y. Soltan-Zadeh, and F. Zayeri, "Global trend of breast cancer mortality rate: a 25-year study," *Asian Pacific journal of cancer prevention: APJCP*, vol. 20, no. 7, p. 2015, 2019.

[2] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling *et al.*, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, 2022.

[3] J. C. Lashof, I. C. Henderson, and S. J. Nass, "Mammography and beyond: developing technologies for the early detection of breast cancer," 2001.

[4] C. D'Orsi, L. Bassett, and S. Feig, "Breast imaging reporting and data system (bi-rads)," *Oxford University Press, New York*, 2018.

[5] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[6] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] I. Hartsock and G. Rasool, "Vision-language models for medical report generation and visual question answering: A review," *Frontiers in Artificial Intelligence*, vol. 7, p. 1430984, 2024.

[8] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[9] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[12] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu *et al.*, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[13] S. Ghosh, C. B. Poynton, S. Visweswaran, and K. Batmanghelich, "Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 632–642.

[14] K. Jain, A. Bansal, K. Rangarajan, and C. Arora, "Mmbcd: Multimodal breast cancer detection from mammograms with clinical history," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 144–154.

[15] L. V. de Moura, R. Ravazio, C. Mattjie, L. S. Kupssinskü, C. M. D. S. Freitas, and R. C. Barros, "Unlocking the potential of vision-language models for mammography analysis," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–4.

[16] A. Urooj Khan, J. Garrett, T. Bradshaw, L. Salkowski, J. Jeong, A. Tariq, and I. Banerjee, "Knowledge-grounded adaptation strategy for vision-language models: Building a unique case-set for screening mammograms for residents training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 587–598.

[17] Z. Cao, Z. Deng, J. Ma, J. Hu, and L. Ma, "Mammovlm: A generative large vision-language model for mammography-related diagnostic assistance," *Information Fusion*, p. 102998, 2025.

[18] P. Oza, U. Oza, R. Oza, P. Sharma, S. Patel, P. Kumar, and B. Gohel, "Digital mammography dataset for breast cancer diagnosis research (dmid) with breast mass segmentation analysis," *Biomedical Engineering Letters*, vol. 14, no. 2, pp. 317–330, 2024.

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[21] H. Thakkar and A. Manimaran, "Comprehensive examination of instruction-based language models: A comparative analysis of mistral-7b and llama-2-7b," in *2023 International Conference on Emerging Research in Computational Science (ICERCS)*. IEEE, 2023, pp. 1–6.

[22] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[23] J. D. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang, "Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–21.

[24] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[25] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[27] M. Renze, "The effect of sampling temperature on problem solving in large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 7346–7356.

[28] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[29] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10 088–10 115, 2023.

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

[31] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11 809–11 822, 2023.

[32] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.