

Comparison of performance of an LLM Council of LFM2.5-1.2B, Qwen3-4B-2507 and Phi-4-Mini-reasoning and Meta LLaMa 3.1 8B against MedGemma 4B on extraction and classification of free-text radiology reports

Vihaan Mathur

Heritage International Xperiential School

Bharat Aggarwal

bharat.aggarwal@maxhealthcare.com

Max Healthcare

Aniket Aman

Max Healthcare

Research Article

Keywords: Large Language Models (LLMs), LLM Council, Radiological Report Analysis, Information Extraction, General-Purpose Reasoning

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11327243/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Comparison of performance of an LLM Council of LFM2.5-1.2B, Qwen3-4B-2507 and Phi-4-Mini-reasoning and Meta LLaMa 3.1 8B against MedGemma 4B on extraction and classification of free-text radiology reports

Vihaan Mathur¹, Bharat Aggarwal²(Corresponding Author), Aniket Aman²

¹Heritage International Xperiential School, Gurgaon

²Max Super Specialty Hospital, West Block, 1 Press Enclave Road, Saket, New Delhi 110017

Abstract

Large language models (LLMs) have demonstrated growing potential for clinical natural language processing (NLP) [1-4], including radiological report analysis, while recent studies have explored both domain-specialised medical models and multi-LLM approaches [7,8] for improving reliability and reasoning. However, the comparative performance of LLM Councils and specialised medical LLMs on structured radiological tasks remains underexplored. This study evaluated an LLM Council framework against MedGemma 4B for radiological report analysis. The Council comprised Qwen3-4B-2507, LFM2.5-1.2B, and Phi-4 Mini Reasoning, with LLaMa 3.1 8B as chairman. We hypothesised that MedGemma would outperform the Council due to its specialised medical training. A total of 783 anonymised mammography, hip-knee-ankle (HKA), and brain magnetic resonance (MR) reports from Max Healthcare Saket were analysed for BI-RADS and ACR extraction, pre- versus post-operative HKA classification, and normal versus abnormal brain MR classification. Performance was assessed using accuracy, F1 score, and task-specific metrics.

MedGemma achieved higher overall accuracy than the Council (88.72% vs. 83.56%; weighted F1: 0.8653 vs. 0.7389), with significant differences overall ($p = 0.0129$) and for HKA classification ($p = 0.0011$), but not mammography or brain MR. However, Qwen3-4B-2507 independently achieved the highest accuracy across all tasks: 91.89% in mammography, 92.93% in HKA, and 90.25% in brain MR, yielding 91.69% overall accuracy and a weighted F1 of 0.8902. Notably, the Council achieved 72.22% HKA accuracy compared with 85.86% for MedGemma. These findings suggest that strong general-purpose reasoning may outperform both consensus-based and specialised approaches for clinical NLP under the conditions evaluated.

Keywords: Large Language Models (LLMs), LLM Council, Radiological Report Analysis, Information Extraction, General-Purpose Reasoning

Acknowledgements

The authors gratefully acknowledge Max Healthcare Institute Ltd for providing access to the anonymised radiological reports and computational workstation required to conduct this study. The authors particularly thank the radiology department at Max Healthcare Saket for facilitating access to the dataset and supporting the computational requirements of the research.

The LLM Council architecture investigated in this study was inspired by the framework proposed by Andrej Karpathy. The open-source repository associated with this framework was cloned and used as the architectural starting point for the experimental implementation. The repository was subsequently adapted for the models, prompts, orchestration, and experimental requirements of this study, and has been listed in the References section.

Table of Contents

Abstract

Acknowledgements

1.0 Literature Review

1.1 Large Language Models in Radiology

1.2 Existing Evaluation Paradigms

1.3 Beyond Single-Model Evaluation

1.4 Research Gap

2.0 Methodology

2.1 Dataset Description

2.2 Preprocessing Overview and Category-Specific

2.3 Final Dataset

2.4 Language Model Selection

2.5 Technical Stack and Experimental Setup

2.6 Experimental Tasks

2.7 Evaluation Metrics

3.0 Results

3.1 Mammography Results

3.2 HKA Results

3.3 Brain MR Results

3.4 Overall Consensus

3.5 Majority Voting Test

3.6 Ablation Test

3.7 Error Analysis

3.8 Computational Resource Analysis

4.0 Discussion

5.0 Limitations and Future Work

6.0 Conclusion

7.0 Appendix

7.1 Class Distribution Tables

7.2 P-Values Table

7.3 Confusion Matrices

7.4 Prompts

7.5 Scripts Used

8.0 Declarations

9.0 References

1.0 Literature Review

1.1 Large Language Models in Radiology

Large language models (LLMs) have generated a considerable level of interest for their applications within healthcare settings, particularly in the field of radiology. The ability of modern LLMs to interpret free-text clinical reports, extract information, perform classification tasks and justify decision-making using natural language instructions provides them with a distinct advantage over traditional machine learning systems which rely on task-specific training [7]. Due to this, LLMs are increasingly being evaluated as tools to automate radiology workflows and reduce administrative burdens which are associated with report interpretation.

There exist multiple studies available which demonstrate high levels of performance in radiology-based NLP tasks using LLMs. These tasks include the extraction of information, report classification and the generation of radiological summaries [1-6]. More recent studies have also reported performance which nears that of expert radiologists in selected text-based tasks, which suggests that LLMs may be able to serve as effective assistants for information extraction and workflow optimisation [1, 2, 6]. Despite this, it should be noted that performance can vary greatly between NLP tasks, the models used, and the evaluation framework tested, which highlights the need for a systemic assessment of the different approaches.

1.2 Existing Evaluation Paradigms

As the usage of LLMs in healthcare settings increases, there are two distinct strategies of development that have been identified in current studies.

The first such strategy focuses on the use of general-purpose models trained on large and diverse datasets, which span multiple domains and are not specialised in any particular field. Models such as GPT, LLaMa, and Qwen frequently demonstrate strong general reasoning capabilities, ability to follow instructions and adaptability over a broad range of tasks [2, 3].

The second such strategy involves the development of medically-specialised models, which are trained on domain-specific datasets with the expectation that the knowledge bases will improve their performance in clinical applications. To achieve this, lineages such as the MedGemma family have been developed through training on biomedical literature, clinical documentation and datasets specific to healthcare and biotech.

However, current studies which compare these approaches have reported mixed findings. While medically specialised models do frequently display advantages in tasks that require more detailed biomedical knowledge, it is still the general-purpose models which frequently outperform medical models in tasks involving the extraction of information, report classification, and structured reasoning [1-4]. These findings suggest that successful performance in clinical NLP tasks may not only require biomedical and healthcare-specific knowledge, but may also require broader reasoning capabilities and instruction-following behaviour.

1.3 Beyond Single-Model Evaluation

Studies which evaluate individual LLMs operating independently in radiology have proven valuable in establishing the baseline capabilities and in identifying performance differences between various systems [1-6]. However, single-model approaches still remain vulnerable to a broad range of limitations, such as hallucination, reasoning errors, and prompt sensitivity, which can heavily affect reliability in clinical settings.

To tackle these limitations, the focus of researchers is shifting more towards the exploration of ensemble and multi-agent approaches that combine outputs from different models. These systems aim to improve their robustness through methods such as consensus, cross-validation and collective reasoning between models. The primary assumption of these systems is that combining multiple perspectives can reduce the weaknesses exhibited by individual models, and thus may produce more reliable outputs [8].

Researchers have explored a wide variety of multi-agent and ensemble architectures to combine the outputs of multiple LLMs. These include debate-based systems in which agents critique and provide feedback to one another's responses, supervisor-based

frameworks that coordinate specialised agents assigned to different subtasks, and consensus-generated approaches which synthesise multiple independent model outputs into one single response [9-11]. Although these architectures vary substantially in their implementation, they share the common aim of improving reliability and reducing errors that may arise from individual model weaknesses.

One implementation of this concept is the ConSensus framework, which generates independent responses from multiple LLMs before producing a synthesised consensus answer through an additional LLM responsible for the final aggregation [11]. Similar supervisor-based architectures employ a ‘coordinating agent’ that integrates information from multiple sources to generate a final decision [10].

The LLM Council architecture which is evaluated in this study most closely resembles a consensus-based architecture. In these systems, multiple LLMs are able to generate responses independently, evaluate the outputs of other LLMs, and then create one final synthesised output produced by a separate decision-making model. Although such architectures have demonstrated promising general reasoning and decision-making task performance, there is still limited evidence regarding their effectiveness and efficiency in clinical and radiological applications.

1.4 Research Gap

Despite the increase in interest in both medically-trained LLMs and multi-agent systems, there is still little evidence currently regarding their performance in radiology report interpretation. Additionally, it is still unclear whether combining multiple models provides any meaningful or concrete advantages over the selection of just a single high-performing model. This distinction has large-scale practical implications in healthcare and clinical settings. Multi-model systems require much greater computational resources, longer inference times, and more complex pipelines than single-model alternatives. If consensus-based architectures do not provide any measurable performance advantages, their additional complexity may not exactly be justified in real-world clinical environments. Thus, in order to address the prevalent research gap, this study will compare the performance of a multi-model LLM Council framework against MedGemma 4B across

three radiological report analysis tasks. Additionally, the study evaluates the performance of the individual constituent Council members to determine whether a consensus-based synthesis offers advantages beyond those achieved by a strong standalone model.

2.0 Methodology

2.1 Dataset Description

A dataset of 800 total free-text radiology reports was provided by Max Healthcare Saket for this study, containing 200 mammograms, 198 HKA reports and 400 Brain MR reports, all prior to preprocessing. The reports range from 2017 to 2025. These reports were randomly sampled and, to align with ethical guidelines, the personally identifiable information of the patients was removed prior to the study, with the report ID being the only identifier in the dataset.

2.2 Preprocessing Overview and Category-Specific

The process of standardising and preprocessing reports was done selectively depending on the experimental task requirements. Certain report categories had various inconsistencies such as incomplete text and variations in their notation, which ultimately necessitated filtering and normalisation prior to the evaluation process.

The inclusion criteria for the mammography reports were the presence of a BI-RADS and ACR score. Reports that failed to meet the criteria were excluded from the study. Out of the 200 initial reports, 15 met the exclusion criteria, resulting in a final dataset of 185 mammography reports. These reports further underwent standardisation to normalise inconsistencies in notation, including Roman numerals and capitalisation, such that BI-RADS scores were represented numerically and ACR scores alphabetically.

No preprocessing or standardisation was performed on the HKA reports due to the task being purely classification-based, and thus the reports were retained in their

original form.

No preprocessing or standardization was done on the Brain MR reports due to the task being purely classification based, and thus the reports were retained in their original form.

2.3 Final Dataset

In total, after preprocessing and standardisation, a dataset of 783 reports was used for this study.

Class distributions were examined for all evaluated tasks to assess potential imbalance within the dataset. HKA classification consisted of pre-operative and post-operative studies, while Brain MR classification consisted of normal and abnormal examinations. Mammography report analysis involved the extraction of Left BI-RADS, Right BI-RADS, and ACR categories, which resulted in multiple class distributions.

A summary of the major class distributions is presented in the Appendix. Although moderate imbalance was present in several tasks, particularly within individual BI-RADS categories, model performance was evaluated using accuracy, precision, recall, F1 score, sensitivity, and specificity where appropriate to reduce reliance on accuracy alone.

Although moderate imbalance was present in several tasks, particularly within the individual BI-RADS categories, model performance was evaluated using accuracy, precision, recall, sensitivity, specificity and F1 score where appropriate to reduce the reliance on accuracy alone. The complete class distribution table is presented in the Appendix.

A small number of mammography reports contained no explicitly documented BI-RADS category and were therefore assigned a null ground-truth label. These cases were retained to reflect real-world reporting variability.

This study utilised retrospective radiological reports obtained from Max Healthcare Saket. All reports were anonymised prior to analysis, with patient identifiers removed

before dataset creation. No patient contact occurred and no identifiable patient information was accessible during the study. The language models were evaluated solely on anonymised report text, and all analyses were performed on de-identified data.

2.4 Language Model Selection

The LLM Council framework implemented for this study utilised locally deployed language models rather than API-based inference systems. This architectural approach was selected to improve inference stability, ensure consistent model accessibility during evaluation, and support a computationally self-contained experimental environment. Consequently, the framework was adapted toward smaller open-source models that remained suitable for local deployment while preserving architectural diversity and instruction-following capability.

This study utilised LFM2.5-1.2B, Qwen3-4B-2507 and Phi-4-mini-reasoning as the base models due to their architectural diversity, reasoning capability, and instruction-following performance. Their comparatively smaller parameter sizes also made them suitable for the locally deployed inference framework utilised in this study.

This study utilised Meta LLaMa 3.1 8B as the chairman model due to its strong performance in prior medical-comparison studies involving radiological and healthcare-related tasks while remaining computationally feasible for the local deployment framework being used.

This study utilised MedGemma 4B independently of the LLM Council as a standalone model due to it being medically-trained and specifically designed for biomedical applications. MedGemma was selected as the benchmark model because it was specifically developed for biomedical and healthcare applications through domain-specific training, making it an appropriate representative of medically specialised language models.

2.5 Technical Stack and Experimental Setup

This study involved the use of a local deployment framework over an API inference system to preserve the methodological structure of the Council framework while supporting computationally feasible local execution. A custom orchestration pipeline was implemented in Python to manage report input, model inference, response collection, and chairman-based synthesis.

The models used were hosted locally on the LM Studio application. For each task, the report dataset was processed through the Council framework by all three base models simultaneously. The generated responses were then anonymised and ranked independently by each model and then the responses with their aggregated rankings were passed to the chairman model for final synthesis.

The experimental implementation was adapted from Karpathy's open-source LLM Council repository [12], with modifications to the model configuration and orchestration to accommodate the requirements of the present study.

The framework was executed on a locally hosted Ubuntu Linux workstation equipped with an AMD Ryzen Threadripper processor, an NVIDIA RTX A4000 GPU, and 128 GB of system memory. All models were deployed locally through the LM Studio inference environment using GGUF quantized variants in the Q4_K_M format to reduce computational overhead while maintaining inference stability and response quality throughout the experimental process.

To ensure consistency and reduce bias during the study, the Council framework was divided into 3 stages. Stage 1 involved providing the same prompt to all base models and the models generating independent responses. Stage 2 involved the models ranking the anonymised outputs to reduce model-specific bias. Stage 3 involved passing on the results of Stages 1 and 2 to the chairman model for final synthesis.

A separate inference pipeline was implemented for MedGemma 4B for direct report-to-model inference. To ensure fairness, MedGemma received the same datasets, task prompts and evaluation conditions. As MedGemma was not part of the Council framework, it did not receive any of the responses, response rankings or the chairman

synthesis.

Although the original LLM Council architecture employs a three-stage consensus process which involves independent model responses, intermediate aggregation and ranking, and finally a chairman-based synthesis, this study utilised a modified version of the architecture. During the preliminary experimentation, the original intermediate aggregation stage frequently led to inference instability, including invalid or even incomplete model outputs when executed within the locally deployed environment. To improve the reliability while preserving the core idea of consensus-based reasoning, the intermediate aggregation stage was replaced with a majority-vote mechanism which was also evaluated. In the implemented workflow, each base model independently generated a prediction, the majority prediction was identified, and these model outputs were subsequently provided to the chairman model for final synthesis. This modification maintained multi-model consensus while ensuring stable execution throughout the experimental evaluation.

2.6 Experimental Tasks

For this study, each report category corresponded to a different experimental task. These tasks evaluated a range of model capabilities, including reasoning, textual interpretation, and instruction-following.

For mammography reports, the models had to extract and output the BI-RADS and ACR scores mentioned. Both standardised reporting systems used in breast imaging assessment. This task evaluated the structured data extraction, instruction-following and interpretation capabilities of the models.

For HKA reports, the models had to classify whether the report indicated a post-operative or pre-operative study. Such classification may assist in determining subsequent clinical management. This task evaluated the reasoning, contextual interpretation, and medical classification capabilities of the models.

For brain MR reports, the models had to classify whether the report indicated a normal or an abnormal study. Such classification may assist in prioritising further clinical

evaluation. This task evaluated the reasoning, contextual interpretation, and medical classification capabilities of the models.

2.7 Evaluation Metrics

Each task had its own set of evaluation metrics to ensure fair comparison between the Council framework and MedGemma and as each task involved different output structures and evaluation requirements.

For mammography reports, the models were evaluated on overall extraction correctness with empirical accuracy, while precision and recall measured the reliability and completeness of extracted values. The F1 score was used to balance precision and recall. The confusion matrix was chosen to enable analysis of category-wise misclassification trends between extracted and ground truth scores.

For HKA reports, the task involved classification of reports into pre-operative or post-operative categories. Empirical accuracy measured overall classification performance, sensitivity and specificity evaluated the ability of the models to correctly identify each class, and the F1 score balanced classification precision and recall. Confusion matrices were used to analyse false positive and false negative classification behaviour.

For Brain MR reports, the task involved classification of reports as normal or abnormal studies. Empirical accuracy evaluated the overall classification correctness, while sensitivity and specificity assessed the models' abilities to correctly identify the abnormal and normal reports respectively. The F1 Score provided balanced evaluation under possible class imbalance, and the confusion matrices enabled a detailed examination of classification errors.

Instances in which a model failed to generate a valid output were recorded as null outputs. For all performance metrics, null outputs were treated as incorrect predictions as failure to generate an interpretable response would represent a practical limitation in real-world deployment. For McNemar statistical testing, the null outputs were excluded to ensure comparison was performed only on cases where both systems produced valid predictions.

McNemar’s test was selected to compare the performance of MedGemma 4B and the Council framework as both systems were evaluated on the same set of radiology reports. Unlike tests which compare independent samples, the McNemar’s test is specifically designed for paired nominal outcomes and evaluates whether two classifiers differ significantly in their pattern of correct and incorrect predictions on the same instances, with significance typically assumed as when $p < 0.05$. The comparison between the two frameworks and Qwen3-4B-2507 was also done to examine the statistical significance between the performances of the model and the frameworks.

This makes it particularly suitable for assessing whether observed performance differences between the two systems are statistically significant rather than arising from random variation.

3.0 Results

3.1 Mammography Results

Table 1: Empirical Accuracy of Left BI-RADS, Right BI-RADS and ACR Extraction

Model	LFM	Phi	Qwen	Council	MedGemma
Left BI-RADS	37.30% 69/185	32.97% 61/185	90.81% 168/185	91.35% 169/185	89.73% 166/185
Right BI-RADS	31.89% 59/185	31.89% 59/185	94.05% 174/184	92.97% 172/185	91.89% 170/185
ACR	17.84% 33/185	18.92% 35/185	90.81% 168/185	87.03% 161/185	90.81% 168/185

Table 2: Average Empirical Accuracy of the models in the mammography tasks

Model	LFM	Phi	Qwen	Council	MedGemma
Average Accuracy	29.01%	27.93%	91.89%	90.45%	90.81%

Table 3: Precision of Left BI-RADS, Right BI-RADS and ACR Extraction

Model	LFM	Phi	Qwen	Council	MedGemma
Left BI-RADS	0.2130	0.8471	0.9349	0.9227	0.9165
Right BI-RADS	0.3888	0.8533	0.9694	0.9455	0.9368
ACR	0.6989	0.6275	0.9452	0.9001	0.9316

Table 4: Recall of Left BI-RADS, Right BI-RADS and ACR Extraction

Model	LFM	Phi	Qwen	Council	MedGemma
Left BI-RADS	0.3750	0.7011	0.9333	0.9185	0.9121
Right BI-RADS	0.3207	0.6782	0.9667	0.9348	0.9341
ACR	0.4286	0.5000	0.9385	0.8750	0.9231

Table 5: F1 Scores of Left BI-RADS, Right BI-RADS and ACR Extraction

Model	LFM	Phi	Qwen	Council	MedGemma
Left BI-RADS	0.2659	0.7083	0.9322	0.9173	0.9085
Right BI-RADS	0.2351	0.6940	0.9664	0.9354	0.9324
ACR	0.2843	0.3925	0.9374	0.8741	0.9214

The confusion matrices of each model on each task are present in the provided Appendix.

As seen in Tables 1-5, Qwen3-4B-2507 achieved the highest empirical accuracy in the Right BI-RADS extraction and shared highest accuracy in the ACR extraction with MedGemma 4B. The Council managed to outperform MedGemma 4B in Left and Right BI-RADS extraction but only outperformed Qwen in the Left BI-RADS extraction.

3.2 HKA Results

Table 6: Empirical Accuracy of HKA Classification

Model	LFM	Phi	Qwen	Council	MedGemma
Empirical Accuracy	76.77% 152/198	32.83% 65/198	92.93% 184/198	72.22% 143/198	85.86% 170/198

Table 7: Sensitivity of HKA Classification

Model	LFM	Phi	Qwen	Council	MedGemma
Sensitivity	0.0488	0.6842	0.8537	0.5366	0.7561

Table 8: Specificity of HKA Classification

Model	LFM	Phi	Qwen	Council	MedGemma
Specificity	0.9554	0.2766	0.949	0.7707	0.9329

Table 9: F1 Scores of HKA Classification

Model	LFM	Phi	Qwen	Council	MedGemma
F1 Score	0.0800	0.3133	0.8333	0.4444	0.7561

The confusion matrices of each model on each task are present in the provided Appendix.

As seen in Tables 6-9, Qwen3-4B-2507 achieved the highest performance across all models evaluated and surpassed the Council performance by over 20%, with MedGemma achieving the second highest accuracy and the Council achieving its lowest accuracy throughout the experiment, significantly bringing down its overall empirical accuracy.

3.3 Brain MR

Table 10: Empirical Accuracy of Brain MR Classification

Model	LFM	Phi	Qwen	Council	MedGemma
Empirical Accuracy	74.50% 298/400	55.25% 221/400	90.25% 361/400	88.00% 352/400	89.50% 358/400

Table 11: Sensitivity of Brain MR Classification

Model	LFM	Phi	Qwen	Council	MedGemma
Sensitivity	0.505	0.700	0.805	0.760	0.9747

Table 12: Specificity of Brain MR classification

Model	LFM	Phi	Qwen	Council	MedGemma
Specificity	0.9850	0.8056	1.0000	1.0000	0.8505

Table 13: F1 Scores of Brain MR classification

Model	LFM	Phi	Qwen	Council	MedGemma
F1 Score	0.6645	0.7420	0.8920	0.8636	0.9190

The confusion matrices of each model on each task are present in the provided Appendix.

As seen in Tables 10-13, Qwen3-4B-2507 achieved the highest empirical accuracy, with MedGemma following close behind by 0.75% and the Council with 2.25%. Qwen and the Council also scored specificities of 1.0000 in the Brain MR classification.

3.4 Overall Consensus

Table 14: Average Accuracies of each task for each model

Model	LFM	Phi	Qwen	Council	MedGemma
-------	-----	-----	------	---------	----------

Average Accuracy	60.09%	38.67%	91.69%	83.56%	88.72%
------------------	--------	--------	--------	--------	--------

Fig 1: Average Accuracy Across All Evaluated Tasks

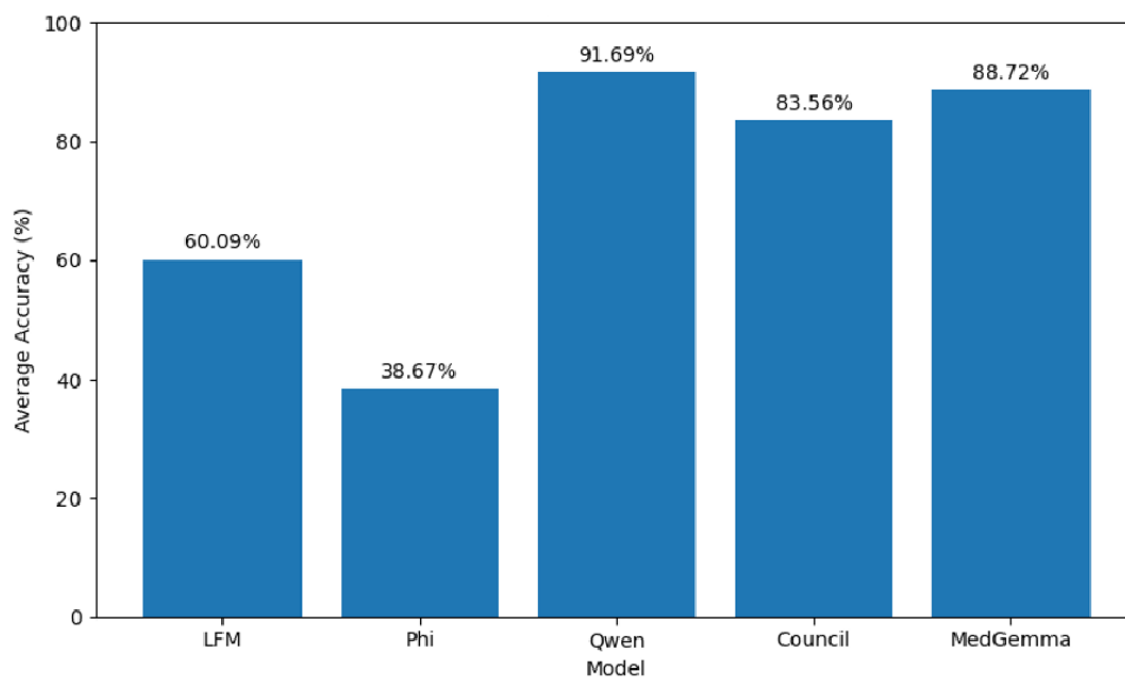


Table 15: Model Accuracy and Confidence Intervals

Model	Pooled Accuracy (%)	95% Confidence Interval
Qwen3-4B-2507	91.50	89.75-92.98
MedGemma 4B	89.51	87.60-91.15

Council	86.47	84.37-88.32
LFM2.5-1.2B	52.99	50.11-55.86
Phi-4-Mini-Reasoning	38.25	35.49-41.09

As seen in Tables 14 and 15 and Fig 1, Qwen3-4B-2507 achieved the highest pooled accuracy (91.50%, 95% CI 89.75–92.98), followed by MedGemma 4B (89.51%, 95% CI 87.60–91.15) and the Council framework (86.47%, 95% CI 84.37–88.32). Confidence intervals for Qwen3-4B-2507 and MedGemma 4B overlapped, suggesting comparable overall performance estimates despite Qwen achieving the highest observed accuracy.

Table 16: P-value between the results of the Council and MedGemma for each task

Task	P-Value as per McNemar Test
Left BI-RADS score extraction	0.605576616335346
Right BI-RADS score extraction	0.789268026134281
ACR score extraction	0.121335250358482
Overall mammography task performance	0.880168454906725
Overall HKA classification performance	0.00105403404578479
Overall Brain MR classification performance	0.576150122030579
Overall experimental performance	0.0129069582052627

As seen in Table 16, overall, the difference in performance between the Council and MedGemma 4B were primarily statistically insignificant in the mammography and Brain

MR tasks, however the overall experimental difference is statistically significant ($p < 0.05$), largely due to the p-value from the HKA task performance difference. The p-values of comparison between Qwen and the Council and Qwen and MedGemma are also provided in the Appendix.

To determine whether or not the Council framework and chairman performance was actually necessary, two additional tests were run. The first test calculated the accuracy of majority voting of the council base models and compared it to the chairman output, and the second was an ablation test which only tested LLaMa 3.1 8B in a similar method that MedGemma was evaluated.

3.5 Majority Voting Test

Table 17: Majority Count Accuracy against Chairman Accuracy

Task	Field	Majority Accuracy (%)
Mammography	Left	92.23
Mammography	Right	90.22
Mammography	ACR	77.78
HKA	Classification	88.95
Brain MR	Classification	87.86
Overall Average Accuracy	-	87.55

The results from Table 17 demonstrate that while there was a similar accuracy between the chairman and majority count in the Left and Right BI-RADS, the majority accuracy decreased by 10% compared to the chairman. The HKA majority accuracy, however is

16% higher than the chairman, indicating the chairman assessment may have introduced additional layers of uncertainty and incorrect decisions during its run. The chairman has a slightly higher accuracy on the Brain MR task. The overall average accuracy is 4% higher than the chairman and 1% lower than MedGemma, indicating that majority voting may be a highly competitive framework in itself in the given tasks.

3.6 Ablation Test

Table 18: Performance of LLaMa 3.1 8B in BI-RADS and ACR Extraction

Task	Empirical Accuracy	Precision	Recall	F1 Score
Left BI-RADS	75.14% 139/185	0.894	0.979	0.885
Right BI-RADS	81.08% 150/185	0.932	0.920	0.922
ACR	85.41% 158/185	0.941	0.935	0.934
Mammography Average	80.54%	0.922	0.945	0.914

Table 19: Performance of LLaMa 3.1 8B in HKA and Brain MR Classification

Task	Empirical Accuracy	Sensitivity	Specificity	F1 Score

HKA	31.82% 63/198	0.854	0.183	0.348
Brain MR	51.25% 205/400	0.532	0.583	0.551

Table 20: Average Accuracy of LLaMa 3.1 8B

Model	Overall Average Accuracy
LLaMa 3.1 8B	54.54%

The results from Tables 18-20 demonstrate that LLaMa 3.1 8B was unable to perform at the level it was performing at as the Council chairman, with an overall average accuracy difference of 29%, and thus justifying the use of the council framework to improve overall model performances.

3.7 Error Analysis

The following examples demonstrate situations in which the models provided correct or incorrect responses.

Example 1: HKA

Ground Truth: Pre-Operative (0)

Report Excerpt: (

Investigation : X-Ray Whole Spine AP & Lateral

Results :

There is reversal of lordosis of spine with osteophyte at multiple levels and multilevel disc space reduction.

The height and shape of the lumbar vertebral bodies appear normal.

Pedicles appear normal.

No obvious prevertebral or paravertebral soft tissue abnormality is visualized.

Advise: Clinical correlation.)

Table 21: Model Predictions for Example 1

Model	LFM	Phi	Qwen	Council	MedGemma
Output	0	1	0	1	0

The report described degenerative spinal changes but contained no evidence of previous surgical intervention, implanted hardware, or other post-operative findings. As per Table 21, Qwen, MedGemma, and LFM correctly identified the study as pre-operative. In contrast, Phi and the Council incorrectly classified the report as post-operative. This example illustrates how errors may arise when narrative clinical findings are interpreted without sufficient attention to the absence of explicit surgical indicators. Additionally, the Council output was an incorrect classification despite two models generating the correct answer, which suggests that consensus-based synthesis does not necessarily preserve the strongest available interpretation.

Example 2: HKA

Ground Truth: Post-Operative (1)

Report Excerpt: (X-ray Whole Spine/Full Limb 07-APR-2017:

Bilateral hip joints show mild reduction in joint spaces.

Early osteoarthritic changes seen in left knee joint.

TKR prosthesis on right side.

Bilateral ankle joint appears grossly normal .

Correlate clinically.)

Table 22: Model Predictions for Example 2

Model	LFM	Phi	Qwen	Council	MedGemma
Output	0	1	1	0	0

The report explicitly states the presence of a “TKR prosthesis on right side”, which provides direct evidence of previous surgical intervention and supporting a post-operative classification. As seen in Table 22, Qwen and Phi correctly classified the study, whereas LFM, MedGemma and the Council incorrectly classified the study despite the presence of clear surgical terminology. Notably, the Council failed to preserve the correct interpretation generated by constituent models, instead producing an incorrect final prediction. This example illustrates that even when post-operative states is explicitly stated within the report, consensus-based synthesis may not necessarily improve the classification accuracy and may fail to retain correct interpretations from individual contributors.

Example 3: HKA

Ground Truth: Post-Operative (1)

Report Excerpt: (Investigation: X-Ray Both lower limbs AP view

Results:

Early degenerative osteoarthritic changes are seen in bilateral knee joints with reduction of tibiofemoral joint spaces more along medial aspects on both sides.

Anatomical and mechanical axes drawn in both lower limbs as shown in the skiagram.

Advise: Clinical correlation)

Table 23: Model Predictions for Example 3

Model	LFM	Phi	Qwen	Council	MedGemma
Output	0	1	0	0	0

The report described “degenerative osteoarthritic changes” and lower-limb alignment measurements, but did not contain any explicit reference to a previous surgery, implanted hardware, prostheses, or any other post-operative findings. As seen in Table 23, despite the ground truth classification being post-operative, four of the five evaluated systems incorrectly predicted a pre-operative study, with only Phi generating a correct classification. Unlike previous examples, this error was not confined to any specific model or the Council framework, and instead affected nearly all evaluated systems. This suggests that certain reports may lack the sufficient textual indicators of surgical status, which makes accurate classification difficult even for high-performing models. This example highlights the inherent limitation of report-based classification tasks, where the ground truth labels may depend on a clinical context which is not explicitly documented within the report itself.

Example 4: Brain MR

Ground Truth: Normal (1)

Report Excerpt: (MRI Spine Cervical (Contrast) of 27-FEB-2017:

Result: MR of the cervical spine was performed. Sequential images were obtained in sagittal and axial planes using T1 and T2 weighted images. Post contrast study was also performed.

Loss of cervical lordosis noted.

Multilevel disc desiccation changes with marginal osteophytes (Disc osteophytic complex) noted in cervical spine from C2 to C7 levels.

C2-C3, C3-C4: Loss of posterior concavity of the disc causing thecal sac indentation. No evidence of spinal nerve root compression / spinal canal stenosis noted.

C4-C5, C5-C6, C6-C7: Diffuse disc bulge causing thecal sac compression. Disc seen encroaching over the bilateral neural foramina, however, no evidence of spinal nerve root compression seen. Mild secondary spinal canal stenosis noted at same levels.

C7-D1: Normal disc morphology. No evidence of nerve root compression. No evidence of spinal canal stenosis noted.

Post contrast images do not showing any abnormal enhancing / paraspinal collection. No focal enhancing lesion noted.

Vertebral bodies are normal in height, morphology and signal intensity.

Cervical spinal cord is normal in morphology and signal intensity.

The CVJ and CMJ are normal.

IMPRESSION: MRI findings reveal disc degenerative changes in cervical spine as described.

Please correlate clinically MT/M

MRI Brain (Contrast) of 27-FEB-2017:

Results: Serial sections of the brain were obtained in sagittal, coronal and axial planes using T1, T2, and FLAIR images. Additional post contrast images were obtained in sagittal, axial and coronal planes.

Brain parenchyma shows normal MR morphology and gray-white differentiation. There is no focal parenchymal lesion.

Post contrast scan show no abnormal meningeal / parenchymal enhancement.

The basal cisterns and superficial subarachnoid CSF spaces are normal.

The ventricular system is normal.

The midbrain, pons and medulla are normal.

The cerebellar hemispheres are normal.

Visualized parts of the sella, 5th, 7th and 8th nerve complexes are grossly normal on routine brain imaging.

Major flow voids are present.

IMPRESSION: Normal MR scan of Brain. No focal lesion is seen in the brain parenchyma.

Please correlate clinically MT/M)

Table 24: Model Predictions for Example 4

Model	LFM	Phi	Qwen	Council	MedGemma
Output	0	Null	1	1	0

The report contained extensive cervical spine degenerative findings, including disc

desiccation, osteophytes, disc bulges, and mild spinal canal stenosis. However, the accompanying brain MRI report was explicitly normal, with the final impression stating: "Normal MR scan of Brain. No focal lesion is seen in the brain parenchyma.". As seen in Table 24, Qwen and the Council correctly classified the study as normal, whereas LFM and MedGemma incorrectly classified it as abnormal. Phi failed to generate a valid response and was therefore recorded as a null output.

This example suggests that some models may have been influenced by a clinically significant abnormality defined elsewhere in the report, rather than restricting their interpretation to the findings relevant to the classification task. The error highlights the importance of contextual focus in radiological report analysis, where the presence of pathology in one anatomical region may not necessarily imply an abnormality in the target study. The null output produced by Phi also illustrates that generation failures do represent a distinct source of performance degradation beyond the conventional classification errors.

Example 5: Brain MR

Ground Truth: Normal (1)

Report Excerpt: (MRI Brain (Contrast) of 18-JAN-2017:

Results:

Serial sections of the brain were obtained in sagittal, coronal and axial planes using T1, T2, and FLAIR images. Additional post contrast images were obtained in sagittal, axial and coronal planes.

Brain parenchyma shows normal MR morphology and gray-white differentiation. There is no focal parenchymal lesion. Post contrast scan shows no abnormal meningeal / parenchymal enhancement.

The basal cisterns and superficial subarachnoid CSF spaces are normal.

The ventricular system is normal.

Bilateral basal ganglia and thalami are normal. No shift of midline structures seen.

The midbrain, pons and medulla are normal. The cerebellar hemispheres are normal.

Visualized parts of the sella, 5th, 7th and 8th nerve complexes are grossly normal on routine brain imaging.

Major flow voids are present, right transverse and sigmoid sinus, and right IJV is mildly hypoplastic.

IMPRESSION: CEMRI study reveals no significant abnormality in brain parenchyma.

Additional finding: Bilateral maxillary polyps with right sided concha bullosa.)

Table 25: Model Predictions for Example 5

Model	LFM	Phi	Qwen	Council	MedGemma
Output	0	0	1	0	1

The report repeatedly described normal brain findings and concluded with the impression that there was "no significant abnormality in brain parenchyma." Although an additional finding of bilateral maxillary polyps and right-sided concha bullosa was noted, these findings were extracranial and did not constitute an abnormal brain MRI. As seen in Table 25, Qwen and MedGemma correctly classified the study as normal, whereas LFM, Phi, and the Council incorrectly classified it as abnormal.

The above example highlights the importance of distinguishing between clinically relevant brain abnormalities and incidental findings which fall outside the target classification task. The Council having an incorrect prediction is notable as it fails despite the presence of a clear-stated normal impression within the report. This suggests that consensus-based aggregation may occasionally amplify the incorrect

interpretations from its constituent models, which results in reduced performance compared to stronger individual systems. The example also supports the observation that high-performing models are generally better able to focus on task-relevant findings, while disregarding any incidental abnormalities which are unrelated to the classification objective.

Example 6: Mammography

Ground Truths: Left BI-RADS = 2, Right BI-RADS = 1, ACR = A

Report Excerpt: (Mammography Both Breast of 02-JAN-2017:

Results:

Mammography using low radiation dose and compression was carried out using a dedicated CR based digital system, in the cranio-caudal and medio-lateral oblique projections.

Both breasts are composed of predominant fat and small fibroglandular parenchyma (ACR A)

Right breast:

No evidence of focal lesion seen.

No evidence of micro or macr O-calcification is seen.

No evidence of skin thickening or nipple retraction is seen.

Right axilla is clear.

Left breast:

A well defined rounded mass seen in MLO view superiorly - ? node.

No evidence of micro or macr O-calcification is seen.

No evidence of skin thickening or nipple retraction is seen.

Left axilla is clear.

Impression: Right breast - BI-RAD 1 category.

Left breast - BI-RAD 2 category.

Adv.: Routine follow up and clinical examination.

Comparison with previous records

MANAGEMENT OF PALPABLE ABNORMALITY SHOULD BE BASE UPON CLINICAL GROUNDS. (Please bring these reports on next visit. False negative rate of mammography is approximately 10%).)

Table 26: Model Predictions for Example 6

Model	LFM	Phi	Qwen	Council	MedGemma
Output	1, 2, B	2, 2, B	2, 1, A	2, 2, B	2, 1, A

The report explicitly stated an ACR density category of A and concluded with a BI-RADS 1 assessment for the right breast and BI-RADS 2 assessment for the left breast. As seen in Table 26, Qwen and MedGemma correctly extracted all three reporting elements. In contrast, the Council correctly identified the left BI-RADS category but incorrectly classified the right BI-RADS category as 2 instead of 1 and the ACR category as B instead of A. Similar errors were observed in the individual outputs of Phi and LFM.

The above example demonstrates how the mammography report analysis was primarily a structured information extraction task rather than a reasoning-intensive classification task. The Council was able to identify certain report elements correctly, but still did fail

to maintain a consistent extraction performance across all three evaluation categories. Furthermore, the incorrect ACR prediction generated by multiple Council constituent models was present in the final output despite the presence of an explicitly stated ACR score within the given text. This suggests that consensus-based aggregation may not guarantee an improvement in extraction accuracy, and may occasionally even reinforce errors which originate from weaker contributions rather than preserving the correct interpretations from stronger models.

Overall, the error analysis revealed multiple recurring patterns which aid in the explanation of the quantitative performance difference observed between models. HKA classification proved to be the most difficult task as it frequently required contextual interpretation of the narrative clinical descriptions rather than direct extraction of explicitly stated findings. Errors in the HKA task were typically associated with missed surgical indicators, ambiguity within the report text, or incorrect reasoning regarding operative status. In contrast, the Brain MR classification errors commonly arose when models were influenced by incidental findings or pathologies described outside the target anatomical region. Mammography errors were generally associated with structured information extraction failures, despite the presence of explicitly stated BI-RADS and ACR categories. Across all tasks, the Council framework occasionally failed to preserve correct interpretations generated by stronger constituent models, suggesting that consensus-based synthesis may introduce additional opportunities for error when model performance is heterogeneous. Collectively, these findings support the quantitative results and indicate that performance differences were driven not only by domain knowledge, but also by differences in contextual reasoning, information extraction accuracy, and robustness of final decision synthesis.

3.8 Computational Resource Analysis

Table 27: Computational Cost Comparison between MedGemma and Council framework

System	Total Parameters	LLM Calls per Report
--------	------------------	----------------------

MedGemma	4B	1
Council	16.2B	4
Computational Ratio (Council : MedGemma)	4.05x	4x

Table 28: Total Inference Time Comparison between MedGemma and Council framework

System	Mammography Total Inference Time (s)	HKA Total Inference Time (s)	Brain MR Total Inference Time (s)	Average Total Inference Time (s)
MedGemma	507.9	540.7	1139.4	729.3
Council	718.5	459.5	1342.26	840.1

Table 29: Average Inference Time Comparison between MedGemma and Council framework

System	Mammography Average Inference Time (s)	HKA Average Inference Time (s)	Brain MR Average Inference Time (s)	Averaged Inference Time (s)
MedGemma	2.746	2.731	2.849	2.775
Council	3.884	2.321	3.356	3.187

Table 30: Peak VRAM Comparison between MedGemma and Council framework

System	Mammography Peak VRAM (GB)	HKA Peak VRAM (GB)	Brain MR Peak VRAM (GB)	Average Peak VRAM (GB)
MedGemma	5.389	5.391	5.393	5.391
Council	13.127	13.164	13.178	13.156

Table 31: Computational Efficiency Comparison between MedGemma and Council framework

	Average Total Inference Time	Averaged Inference Time	Average Peak VRAM
Computational Cost Ratio (Council : MedGemma)	1.152x	1.148x	2.440x

As seen in Tables 27-31, overall, the Council framework required substantially greater computational resources than MedGemma 4B. While MedGemma consisted of a single 4B-parameter model requiring one LLM inference call per report, the Council framework incorporated four simultaneously loaded language models with a combined parameter count of 16.2B and required four inference calls per report, representing approximately a 4.05-fold increase in model parameters and a fourfold increase in inference operations. Despite this, the observed computational overhead was comparatively modest. Across all tasks, the Council required an average inference time of 3.187 seconds per report compared with 2.775 seconds for MedGemma, corresponding to a

1.15-fold increase in inference time. Similarly, average total processing time increased from 729.3 seconds to 840.1 seconds (1.15×). The largest increase was observed in GPU memory requirements, where the Council required an average peak VRAM allocation of 13.156 GB compared with 5.391 GB for MedGemma, representing a 2.44-fold increase. These findings demonstrate that despite the Council architecture substantially increasing the computation requirements through additional models and inference stages, the increase in execution time is much smaller than the increase in parameter account. Nevertheless, the higher memory requirements and more complex pipeline did not produce measurable improvements in its predictive performance, indicating that the additional computational cost was not justified under the conditions evaluated in this study.

4.0 Discussion

The main goal of this study was to compare the performance and accuracy of the consensus-based LLM Council framework against MedGemma 4B in radiological report analysis tasks. While MedGemma was able to outperform the Council overall, the most important finding came from an individual Council constituent model. Qwen consistently achieved the highest performance accuracy across the tasks evaluated, outperforming both MedGemma and the Council outputs. This finding suggests that, for the report-analysis tasks evaluated in this study, strong general-purpose reasoning capabilities may be as important as, if not more important than, domain-specific medical training or consensus-based aggregation.

The superior empirical accuracy of Qwen across the tasks is notable as it is a general-purpose LLM without explicit medical training. This does suggest that the evaluated tasks are heavily influenced by structured reasoning above clinical text rather than requiring a deep and detailed domain-specific medical knowledge base alone. In the mammography report analysis, Qwen was able to reliably extract the structured diagnostic scores, while

in the Brain MR classification it was able to accurately distinguished between normal and abnormal findings based on explicit textual cues.

The most notable task was the HKA classification task as it produced the most significant behavioural difference, and is also where Qwen achieved its largest performance margin over both MedGemma and the Council. The task requirements make it more dependent on contextual reasoning and multi-step inference rather than direct extraction. The consistently strong performance of Qwen here suggests that general reasoning ability and narrative understanding may be equally or more important than domain-specific training for certain clinical NLP tasks.

MedGemma 4B did demonstrate a strong and stable performance across the evaluated tasks, especially in HKA classification where it outperformed the Council framework by 13%. Its performance does indicate its robustness in medically-grounded interpretation tasks, which most likely reflects its domain-specific training. However, its inability to consistently outperform Qwen suggests that medical specialisation alone is insufficient when tasks require a combination of structured reasoning and contextual interpretation.

The Council framework results demonstrate task-dependent performance variability rather than uniformly poor performance. The framework achieved competitive results in the mammography and the brain MR tasks, however it significantly underperformed in HKA classification which ultimately drove its lower overall performance. This suggests that the Council framework is more suitable for more structured extraction tasks, but is less reliable in contexts which require a deeper semantic interpretation.

Notably, the Council framework consists of multiple general-purpose models, with LLaMa 3.1 8B acting as chairman for final synthesis. Despite the inclusion of the strongest-performing model, the system did not surpass its best individual component.

This finding suggests that consensus-based aggregation may not necessarily improve performance when substantial differences exist between the constituent models. In these cases, stronger model outputs may get diluted weaker contributions during synthesis, which results in final predictions that reflect agreement over correctness. This highlights

a key limitation of the consensus-based architectures for structured radiological report-analysis tasks.

Task-level analysis further aids in the examination of these differences. Mammography extraction tasks demonstrated lightly clustered performance across the models, which reflects the structured nature of BI-RADS and ACR extraction where answers are often explicitly stated in the report text. Brain MR classification showed moderate variation but stable ranking consistency, which suggests a balance between the explicit cue recognition and limited inference requirements. In contrast, the HKA classification showed the largest divergence between the models, highlighting its reliance on contextual interpretation of clinical narratives and its sensitivity to reasoning inconsistencies and aggregation errors. As shown in the Error Analysis section, multiple HKA misclassifications were associated with difficulties in the interpretation of narrative descriptions of surgical status rather than an explicit procedural terminology.

Overall, these patterns do indicate that model performance is largely dependent on the level of contextual reasoning required. Tasks dominated by structural expansion produce relatively small performance differences between models, while tasks which required multi-step inference amplify differences in reasoning robustness. This finding does explain the superior consistency demonstrated by Qwen and the instability within the Council framework in more inference-heavy settings.

The experimental findings have various implications for medical LLM design. While medically specialised models such as MedGemma 4B itself are expected to perform with a superior accuracy due to domain-specific training, this advantage appears to be task-dependent. In situations where medical terminology gets combined with structured reasoning and information extraction, general-purpose models could achieve comparable or even a superior performance. This suggests that general reasoning abilities and narrative understanding can play a crucial role alongside domain-specific training in clinical NLP tasks.

A particularly notable difference in behaviour was observed in the Brain MR task. MedGemma achieved a substantially higher sensitivity than Qwen, which indicates a stronger tendency to identify abnormal studies. However, this came together with lower specificity, resulting in more false positives. By contrast, Qwen demonstrated perfect specificity while maintaining a high overall accuracy. This suggests that there exist differing classification thresholds between models, with MedGemma favouring sensitivity while Qwen favours conservative but more precise predictions.

Additionally, the high accuracy of the majority voting system does indicate that the inclusion of a chairman model for final synthesis may not be entirely necessary and may even decrease performance as per the conditions of this study. While the ablation test does justify the use of a Council architecture to improve the chairman model's individual performance, its inclusion may have decreased Council performance in the evaluated tasks.

General-purpose models can benefit from broader optimisation across reasoning, abstraction and instruction adherence, which allows for more flexible adaptation to mixed reasoning-extraction tasks. This suggests that optimal model selection would depend not only on domain specificity, but also on balance between reasoning complexity and clinical task structure.

Finally, the performance of the Council architecture highlights the limitations of consensus-based approaches in this setting. Despite having a higher computational cost, parallel inference across multiple models, and the synthesis stage by the chairman model, the Council was unable to provide measurable performance gains over Qwen. Instead, it demonstrated reduced accuracy and F1 performance, particularly in the reasoning-intensive tasks. Under the conditions of this study, these results do definitively suggest that consensus-based multi-model systems may not be ideal for radiological report analysis when compared to stronger single-model alternatives that offer high efficiency and more stable performance.

5.0 Limitations and Future Work

Multiple limitations need to be considered during the interpretations of the findings. The study was conducted with reports procured only from a single institution and focused only on text-based report-analysis tasks. Consequently, the results should not be taken as evidence that general-purpose always outperform medical LLMs across all healthcare applications.

Several directions for future research emerge from these limitations.

One extension is to evaluate the Council and individual models on larger and multi-institutional datasets. This would allow an assessment across diverse clinical settings and reduce biases that may be introduced by institution-specific reporting styles.

Another such extension could explore more systematic ablation studies on the Council's architecture. This extension could evaluate the impact of different chairman models, alternative aggregation strategies such as weighted voting, and the removal of weaker constituent models. Such an analysis would significantly aid in determining whether performance limitations arise from the consensus mechanism used or from the specific configuration of the Council used in this study.

Furthermore, systematic ablation studies on the Council architecture could be conducted to examine how different configurations, chairman models, and consensus methods can influence the architecture's performance. This analysis would aid in determining whether performance is limited by the consensus mechanism or from a specific configuration.

Furthermore, expanding the evaluation datasets to incorporate multimodal inputs, specifically radiological imaging with reports, would also provide a more practical assessment of clinical reasoning capabilities. This examination would better reflect real-world radiology workflows where the dual-incorporation of text and images in reports is essential.

Another promising direction involves the usage of retrieval-augmented generation (RAG) pipelines, which would allow the models to access structured medical knowledge bases or radiological guidelines during inference. This may improve performance in tasks which require domain-specific knowledge beyond that which is present in the report alone.

Finally, future work could investigate hybrid systems combining strong general-purpose models with medical domain-specific models in adaptive or dynamically weighted ensembles. Given the strong performance of Qwen3-4B-2507 observed in this study, such hybrid architectures may provide a balance between reasoning capability and domain expertise.

6.0 Conclusion

This study evaluated the performance of a consensus-based LLM Council framework against MedGemma 4B, a medically specialised language model, across three radiological report analysis tasks: mammography BI-RADS/ACR extraction, hip-knee-ankle (HKA) classification, and brain MR classification. The objective was to determine whether a multi-model architecture comprising general-purpose language models and chairman-based synthesis could match or exceed the performance of a domain-specific medical model in structured clinical NLP tasks.

Overall, MedGemma 4B outperformed the Council framework, particularly in the HKA classification task, where contextual interpretation and clinical reasoning played a greater role than direct information extraction. Whilst the Council achieved competitive performance in structured extraction tasks, its performance declined in reasoning-intensive scenarios, resulting in lower overall accuracy and F1 scores.

The most important finding, however, emerged from the performance of an individual Council member. Across all evaluated tasks, Qwen3-4B-2507 consistently achieved the highest overall performance, surpassing both MedGemma 4B and the final Council

output. This suggests that strong reasoning ability, instruction-following behaviour, and robust language understanding may be as important as, and in some settings more important than, domain-specific medical training for report-analysis tasks.

Despite requiring additional inference stages, much greater computational resources and increased implementation complexity in the pipelines, the Council framework did not outperform its strongest constituent model. Under the conditions evaluated in this study, aggregating multiple model outputs appear to dilute rather than enhance the Council performance, particularly in tasks which require contextual reasoning.

Altogether, these results suggest that increasing the architectural complexity does not necessarily lead to a better clinical NLP performance. For radiological report-analysis tasks similar to those evaluated in this study, a strong standalone model may be more efficient and a more effective solution than either a specialised medical model or a multi-model consensus framework. Future studies involving larger multi-institutional datasets, alternative Council configurations, and multimodal imaging-report workflows will be important to determine the level of generalisation of these findings and to better understand when, and if, consensus-based aggregation approaches offer meaningful advantages in clinical AI applications.

7.0 Appendix

Provided as an additional file. Tables 32-35 depict class distribution, and Table 36 depicts the p-value results. The confusion matrices of each model for each task are shown from Fig 2-26. The appendix also contains the exact prompts used for the Council, MedGemma, and LLaMa ablation test. It also contains the link to the GitHub repository which contains all the scripts and graphs used during the experimentation.

8.0 Declarations

Funding Declaration: This research project has received no funding from any funding agency in the public,

commercial, or not-for-profit sectors.

Clinical Trial Number: Not applicable

Consent to Publish: Not applicable

Ethics Declaration and Consent to Participate: This research project has received the waiver of consent from Max Healthcare Ethics Committee via approval number IEC REF.NO.: BHR/RS/MSSH/MHIL/SKT-1/MHEC/RADIO/26-23 to allow use of retrospective de-identified data.

9.0 References

- [1] B. Le Guellec, A. Lefèvre, C. Geay, L. Shorten, C. Bruge, L. Hacein-Bey, P. Amouyel, J.-P. Pruvo, G. Kuchcinski, A. Hamroun, Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports, *Radiol. Artif. Intell.* 6 (2024) e230364. <https://doi.org/10.1148/ryai.230364>.
- [2] M.M. Kanani, A. Monawer, L. Brown, W.E. King III, Z.D. Miller, N. Venugopal, P.J. Heagerty, J.G. Jarvik, T. Cohen, N.M. Cross, High-Performance Prompting for LLM Extraction of Compression Fracture Findings from Radiology Reports, *J. Imaging Inform. Med.* 39 (2026) 973–988. <https://doi.org/10.1007/s10278-025-01530-6>.
- [3] J. Wihl, E. Rosenkranz, S. Schramm, C. Berberich, M. Griessmair, P. Woźnicki, F. Pinto, S. Ziegelmayer, L.C. Adams, K.K. Bressemer, J.S. Kirschke, C. Zimmer, B. Wiestler, D. Hedderich, S.H. Kim, Data Extraction from Free-Text Stroke CT Reports Using GPT-4o and Llama-3.3-70B: The Impact of Annotation Guidelines, *Eur. Radiol. Exp.* 9 (2025) 61. <https://doi.org/10.1186/s41747-025-00600-2>.

- [4] M.S. Jabal, P. Warman, J. Zhang, K. Gupta, A. Jain, M. Mazurowski, W. Wiggins, K. Magudia, E. Calabrese, Open-Weight Language Models and Retrieval-Augmented Generation for Automated Structured Data Extraction from Diagnostic Reports: Assessment of Approaches and Parameters, *Radiol. Artif. Intell.* 7 (2025) e240551. <https://doi.org/10.1148/ryai.240551>.
- [5] A. Fink, A. Rau, M. Reiser, F. Bamberg, M.F. Russe, Retrieval-Augmented Generation with Large Language Models in Radiology: From Theory to Practice, *Radiol. Artif. Intell.* 7 (2025) e240790. <https://doi.org/10.1148/ryai.240790>.
- [6] Y. Li, R. Gupta, S. Thompson, Enhancing Large Language Models with Retrieval-Augmented Generation: A Radiology-Specific Approach, *Radiol. Artif. Intell.* 7 (2025) e240313. <https://doi.org/10.1148/ryai.240313>.
- [7] I.Y. Chen, P. Rajpurkar, E.J. Topol, The Current and Future State of AI Interpretation of Medical Images, *N. Engl. J. Med.* 390 (2024) 1234–1245. <https://doi.org/10.1056/NEJMra2301725>.
- [8] Meshuggah22, Andrej Karpathy's LLM Council: When Ensemble Learning Meets Large Language Models, Medium, 2024. <https://medium.com/@meshuggah22/andrej-karpathys-llm-council-when-ensemble-learning-meets-large-language-models-e3312fd02064> (accessed 24 June 2026).
- [9] Y. Liang, Y. Wu, S. Wang, Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate, *arXiv preprint* (2023). [arXiv:2305.19118](https://arxiv.org/abs/2305.19118). <https://arxiv.org/abs/2305.19118>.
- [10] Y. Du, S. Li, A. Torralba, J.B. Tenenbaum, I. Mordatch, Improving Factuality and Reasoning in Language Models through Multiagent Debate, *arXiv preprint* (2023). [arXiv:2305.14325](https://arxiv.org/abs/2305.14325). <https://arxiv.org/abs/2305.14325>.
- [11] X. Zhang, H. Chen, Y. Liu, ConSensus: Multi-Agent Collaboration for Multimodal Sensing, *arXiv preprint* (2026). [arXiv:2601.06453](https://arxiv.org/abs/2601.06453). <https://arxiv.org/abs/2601.06453>.

[12] A. Karpathy, LLM Council, GitHub repository. <https://github.com/karpathy/llm-council> (accessed 25 June 2026).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [LLMCouncilAppendixVihaanMathur.docx](#)