Large Language Model Performance in Interpreting Radiographic Reports: Benchmarking Against Expert Interpretations

Cory Clemmons, Laxman Maheshkar, Sharven Rane, Ben Cote

Vetology Innovations LLC

Abstract

This paper investigates the use of large language models in veterinary radiology to classify, summarize, and interpret reports. Datasets on canine thorax, canine abdomen, and feline thorax were used to benchmark performances of GPT-4O Mini and Gemini 1.5 Flash against expert-verified gold standards. Both models performed quite well in classification, and Gemini 1.5 Flash showed a slight edge over two of the three datasets tested. The major metrics taken into consideration are accuracy, sensitivity, specificity, and F1 scores (Table 1).

The designed prompts were very sensitive, with the goal of patient safety through flagging even minor abnormalities for further investigation. Several challenges related to terminology inconsistencies, overlying conditions, and ambiguity in reports were also addressed. The results emphasize the models' potential to support veterinary diagnostics in improving consistency and efficiency and reducing diagnostic oversight.

This project highlights how the interpretation of veterinary radiology should be performed with the help of AI tools, which greatly increases the speed of work, gives a second opinion and become a cost-effective alternative which in turn has the potential to improve the quality of veterinary practice on a large scale.

Introduction

In recent years, the application of AI in the field of healthcare has grown tremendously as many diagnostic processes have been improved through machine learning models especially for humans. In contrast, veterinary practices, especially in radiology, still need to be still be explored in the same fashion about such advances. Radiology is essential for the establishment of various diagnoses in animals, but this field too has its problems, including the scarcity of veterinary radiologists, the inconsistencies in the interpretations of complex cases and they may, however, introduce long delays in establishing the correct diagnosis and even provide some stability in clinical outcomes, which reinforces the need for novel strategies.

An interesting approach to these challenges is provided by LLMs and generative AI, which may assist in analysing and interpreting reports of radiology. Although the application of LLMs as language agents has proved to be successful for many tasks that involve understanding and generating natural language, their utility in veterinary practice has potential yet to be fully unveiled. The contemporary systems hardly provide understanding from a viewpoint of a specific domain, being confined to structured data that is, too, scant in regard to unstructured reports in radiology.

A unique LLM-based model that specifically focuses on classifying, summarizing

and interpreting radiology reports for veterinary patients was developed in this paper. The model AI enables this by utilizing a well-structured radiology finding datasets

1. Classification Challenges

A number of issues impede the challenge of effectively categorising illnesses in radiology reports using large language models (LLMs). One of the most major issues is the overlap of symptoms across various diseases, which makes it difficult to accurately categorise ailments.

LLMs also require particular prompts that can handle a large corpus of information in order to diagnose illnesses effectively. To obtain exact results, skilled specialists must often conduct in-depth analyses of both radiological pictures and reports.

Moreover, obtaining a sufficient number of high-quality radiographs for each condition remains a major barrier, limiting the ability of training models to identify disease accurately.

2.Dataset

Our dataset was composed of veterinary radiology records collected from a reputed company. These reports were divided into three spreadsheets according to species and body region: canine abdomen, canine thorax, and feline thorax. Each report included significant patient-related data for 50 cases, organized into individual spreadsheets.

For our study, we focused on the following important areas.

• **Findings**: The radiologist's key observations from the radiograph.

• **Conclusions**: Radiologist's final diagnosis.

• **Recommendations**: Suggested treatment plans for the patient.

•**Misc.:** Additional information in the report

This categorisation by species and body region increased flexibility in selecting relevant keywords and parameters, allowing for more efficient input processing for the large language model (LLM).

3. Methodology

This approach first imports the radiology dataset from an Excel file. Missing values are replaced with empty strings in the dataset to ensure that the concatenation and processing are done correctly.

Important features of the reports are combined into a single textual format, which would enable the model to analyse them more effectively.

To extract key conditions from reports, a comprehensive prompt was designed for an Open AI GPT-4o mini model. This prompt provides the requirements to be verified, along with categorization rules and conditional dependencies. The model analyses the text and then makes a determination about each condition, whether Normal or Abnormal, and why.

The application uses Python's Open AI API for LLM interactions and Pandas for data processing. The model's replies are processed to obtain clear classifications, along with explanations consistent with specified standards.

3.1 Preprocessing

To ensure the dataset's relevance, all sensitive information, including patient and client identification, was anonymised to respect confidentiality and ethical requirements.

To prepare the dataset for analysis and integration with a large language model (LLM), numerous pre-processing procedures were taken:

1.Text Cleaning:

To standardize the content, extraneous characters were eliminated, such as special symbols and formatting artifacts (e.g., \n, tabs). Spelling mistakes and veterinary-specific acronyms were also standardized.

2.Segmenting Reports:

Each report was divided into several components, including Recommendations, Conclusions, and Findings. To maintain their contextual relevance for the model, these segments were tagged.

3.Handling Missing Data:

To ensure consistent input length, reports with missing or partial portions were either eliminated or replaced with placeholders.

4.Formatting for Input:

The segmented and cleaned text was organized into a prompt format that was suitable with the LLM. Anonymized patient ID fields, report sections, and other case-related metadata were all included in each prompt entry.

These pre-processing procedures made sure the dataset was clear, reliable, and appropriate for assessing and optimizing the LLM in veterinary radiology jobs.

3.2 Model Architecture and Training

Advanced language models used in this study include Gemini Flash 1.5 and GPT 4 o Mini. These models have been chosen because they can process complex text data to make highly accurate predictions for specialized domains such as veterinary radiology.

Gemini Flash 1.5: It has been renowned for its speed in processing, used to manage such elongated radiology reports quickly and with contextual precision. In architecture, this supports fast inference and is good for the extraction of relevant patterns from unstructured data.

GPT 4 o Mini: This compact version of GPT was selected due to its light architecture and the capability to fine-tune very well with smaller-sized datasets. Despite its size, it did quite well in the area of catching veterinary-specific terminology and clinical nuances.

Prompt Engineering: A structured prompt is generated listing conditions to identify in the report, such as pneumonia, cardiomegaly, or diseased lungs.

In the prompt, logical dependencies are embedded to make the consistency of classification regarding related conditions, such as "diseased lungs contain perihilar infiltrate.".

Model Response Processing:

The API facilitates this prompt and provides condition classification as either "Normal" or "Abnormal," along with a rationale for each determination.

The output is extracted using regex to standardize it for integration with the original dataset.

Fine-Tuning and Customization

Although the models were employed without additional fine-tuning, careful crafting of prompts was used as a method to adapt the model responses to veterinary terminology. This includes:

Incorporate clinical rules specific to radiology, such as dependencies between conditions.

Ensuring clarity in distinguishing normal findings that are appropriate to the patient in context, such as age-related or weight-related normalcy.

3.3 Evaluation Metrics & Challenges

In order to evaluate the performance of the LLM-based classification system, the output was stringently validated against an expert-derived gold standard. This gold set consisted of radiologist-verified classifications for the same set of radiology reports, ensuring a high degree of accuracy and relevance. This comparison allowed for the determination of model performance across a range of conditions.

The performance of the LLM model was evaluated using a **confusion matrix**. Each condition was analysed to determine the following metrics:

- **True Positives (TP)**: Conditions correctly identified as "Abnormal" by the model and present in the expert gold set.
- **True Negatives (TN)**: Conditions correctly identified as "Normal" by the model and confirmed as normal in the expert gold set.
- **False Positives (FP)**: Conditions incorrectly classified as "Abnormal" by the model, but considered "Normal" by the expert.
- **False Negatives (FN)**: Conditions incorrectly classified as "Normal" by the model, but identified as "Abnormal" in the expert set.
- **Accuracy**: Indicates the accuracy of predictions with respect to gold set.
- **Recall (Sensitivity):** Measures the model's ability to correctly identify abnormal conditions.
- **Specificity**: Highlights the model's capability to correctly classify normal conditions.
- **F1 Score**: Combines Precision and Recall to provide a balanced performance metric.

To further make it robust, first, a subset of the dataset containing 50 rows was tested with results carefully analysed. This smaller subset allowed confirmation that the prompt and the classification logic functioned as intended before scaling to the complete dataset. Further comparisons across multiple subsets also confirm the consistency of the model.

Challenges and Solutions

The generated classifications and explanations were reviewed by veterinary radiologists for clinical validity. Expert feedback was used in

refining the prompt to improve logical dependencies, for example, ensuring that related conditions like cardiomegaly and left-sided cardiomegaly will be classified consistently.

Overlapping Conditions: Many conditions, such as diseased lungs and pulmonary nodules, showed significant overlap. Custom prompt logic ensured that broader conditions were flagged only if sub-conditions were classified as "Abnormal."

Ambiguity in Text: Some of the reports contained ambiguous language. The prompt was iteratively refined to handle cases where conditions were marked as "normal for the patient's age or weight."

Balancing between False Positives and False Negatives: The model was tuned to prefer reducing False Negatives, because missing an abnormal condition might have serious clinical consequences.

The radiology reports used various ways to describe the same finding, such as "fluid accumulation" and "effusion." This made proper classification difficult. As a result, a standardized dictionary of terms and synonyms was created, along with a prompt refinement, so that conditions were matched to consistent terminology.

4. Results

This paper presents the results of a study investigating the performance of two large language models, GPT-4O Mini and Gemini 1.5 Flash, in classifying radiology conditions for three datasets: canine thorax, canine abdomen, and feline thorax, as either normal or abnormal. It reports aggregated and individual condition-specific metrics in an attempt to present a comprehensive view regarding the effectiveness of these models. Results for accuracy, sensitivity, specificity, and F1-score are reported.

These model prompts have been devised in a very sensitive manner to identify even slight abnormalities because all findings are potentially of significance and may be representative of a situation concerning patient safety. This implies the tendency always to err on the side of caution, providing models helpful for identifying the case requiring further study rather than prematurely closing the patient case as normal.

Complementing the aggregated metrics, detailed results for each dataset are provided in order to further outline the performance of GPT-4O Mini and Gemini 1.5 Flash on condition-specific classifications. These results are shown as confusion matrices that report the TP, TN, FP, and FN for each condition.

Besides, accuracy values are included to give a better perspective on how each model would perform in differentiating between normal and abnormal conditions. The way these metrics are presented juxtaposes the strengths and relative weaknesses of both models regarding different datasets and conditions.

The following tables outline results for each dataset-canine thorax, canine abdomen, and feline thorax-pooled by condition.

Dataset	Model	Accuracy	Sensitivity	Specificity	F1-Score
Canine Abdomen	GPT-4 o Mini	94%	0.93	0.94	0.78
Canine Abdomen	Gemini 1.5 Flash	95.40%	0.78	0.98	0.80
Feline Thorax	GPT-4 o Mini	96.36%	0.96	0.96	0.82
Feline Thorax	Gemini 1.5 Flash	95.81%	0.90	0.96	0.79
Canine Thorax	GPT-4 o Mini	95.70%	0.95	0.96	0.83
Canine Thorax	Gemini 1.5 Flash	95.80%	0.87	0.97	0.82

Table 1: Quick Overview of Key Metrics Across All Three Datasets

	GPT 4 o Mini					Gemini Flash 1.5				
Condition	TP	FN	TN	FP	**Accuracy**	TP	FN	TN	FP	**Accuracy**
Gastritis	7	1	34	8	**0.82**	5	3	41	3	**0.92**
Ascites	3	0	47	0	**1**	3	0	47	0	**1**
Colitis	12	0	32	6	**0.88**	6	6	37	6	**0.86**
Liver mass	1	0	45	4	**0.92**	1	0	46	0	**0.94**
Pancreatitis	10	2	35	3	**0.9**	10	2	36	2	**0.92**
Microhepatia	5	0	45	0	**1**	4	1	45	1	**0.98**
Small Intestinal obstruction	6	0	42	2	**0.96**	6	0	43	0	**0.98**
Splenic Mass	3	1	46	0	**0.98**	3	1	46	0	**0.98**
Splenomegaly	1	0	49	0	**1**	1	0	48	1	**0.98**
Hepatomegaly	6	0	41	3	**0.94**	6	0	43	1	**0.98**

Table 2.1: Canine Abdomen Classification Report for Each Condition

	GPT 4 o Mini					Gemini Flash 1.5				
Condition	TP	FN	TN	FP	**Accuracy**	TP	FN	TN	FP	**Accuracy**
Pulmonary Nodules	7	0	41	2	**0.96**	7	0	43	0	**1**
Esophagitis	2	0	48	0	**1**	1	1	48	0	**0.98**
Pneumonia	1	0	45	4	**0.92**	1	0	46	3	**0.94**
Bronchitis	22	0	28	0	**1**	21	1	28	0	**0.98**
Interstitial	3	1	43	3	**0.92**	3	1	42	4	**0.90**
Diseased Lungs	29	3	15	3	**0.88**	28	4	18	0	**0.92**
Hypo Plastic Trachea	0	0	50	0	**1**	0	0	50	0	**1**
Cardiomegaly	12	0	38	0	**1**	12	0	38	0	**1**
Pleural Effusion	1	0	49	0	**1**	1	0	49	0	**1**
Perihilar Infiltrate	0	0	49	1	**0.98**	0	0	49	1	**0.98**
RTM	0	0	50	0	**1**	0	0	50	0	**1**
Focal Caudodorsal Lung	4	0	40	6	**0.88**	2	2	42	4	**0.88**
Right Sided Cardiomegaly	1	0	45	4	**0.92**	1	0	38	11	**0.78**
Focal Perihilar	0	0	49	1	**0.98**	0	0	49	1	**0.98**
Left Sided Cardiomegaly	1	0	46	3	**0.94**	1	0	41	8	**0.84**
Bronchiectasis	0	0	45	5	**0.9**	0	0	50	0	**1**
Pulmonary Vessel Enlargement	2	0	48	0	**1**	2	0	48	0	**1**
Thoracic Lymphadenopathy	1	0	47	2	**0.96**	1	0	47	2	**0.96**
Pulmonary Hypo inflation	1	0	49	0	**1**	1	0	49	0	**1**
Pericardial Effusion	0	0	50	0	**1**	0	0	50	0	**1**
Fe Alveolar	0	0	50	0	**1**	0	0	49	1	**0.98**

Table 2.2: Feline Thorax Classification Report for Each Condition

	GPT 4 o Mini					Gemini Flash 1.5				
Condition	TP	FN	TN	FP	**Accuracy**	TP	FN	TN	FP	**Accuracy**
Perihilar Infiltrate	4	0	40	6	**0.88**	4	44	2	0	**0.96**
Pneumonia	4	0	40	6	**0.88**	4	43	3	0	**0.94**
Bronchitis	15	2	32	1	**0.94**	11	33	0	6	**0.88**
Interstitial	11	0	36	3	**0.94**	11	38	1	0	**0.98**
Diseased lungs	24	1	24	1	**0.96**	22	25	0	3	**0.94**
Hypo Plastic Trachea	1	0	49	0	**1**	0	48	1	1	**0.96**
Cardiomegaly	9	0	41	0	**1**	9	41	0	0	**1**
Pulmonary Nodules	6	0	43	1	**0.98**	6	43	1	0	**0.98**
Pleural Effusion	2	0	47	1	**0.98**	2	47	1	0	**0.98**
RTM	2	0	48	0	**1**	0	47	1	2	**0.94**
Focal Caudodorsal Lung	6	0	39	5	**0.9**	6	37	7	0	**0.86**
Focal Perihilar	3	0	41	6	**0.88**	3	44	3	0	**0.94**
Pulmonary Hypo Inflation	4	0	44	2	**0.96**	4	45	1	0	**0.98**
Right Sided Cardiomegaly	2	0	45	3	**0.94**	2	44	4	0	**0.92**
Pericardial Effusion	1	0	49	0	**1**	1	49	0	0	**1**
Bronchiectasis	0	0	48	2	**0.96**	0	49	1	0	**0.98**
Pulmonary Vessel Enlargement	1	0	48	1	**0.98**	1	49	0	0	**1**
Left Sided Cardiomegaly	6	0	44	0	**1**	6	42	2	0	**0.96**
Thoracic Lymphadenopathy	1	2	47	0	**0.96**	2	47	0	1	**0.98**
Esophagitis	1	0	49	0	**1**	0	49	0	1	**0.98**

Table 2.3: Canine Thorax Classification Report for Each Condition

The results make it evident that the:

- **Overall Performance**: Both models attained strong classification accuracy on all three datasets, with especially high strengths in the detection of abnormalities across all conditions.
- **Sensitivity to Abnormalities**: High sensitivity values for both models assure the reliability of detection in abnormal conditions, hence supporting the designed prompts for safety in a patient.
- **Variations Across Conditions**: Slight variations were noted in the performance across different conditions and datasets, which indicate those aspects where further fine-tuning or additional training data might help the models.
- **Comparison of Model**: Overall, though both models scored comparably, the Gemini Flash 1.5 model had a slight upper hand in 2 out of 3 datasets.

5. Conclusion

The results demonstrate robust performance across three datasets canine thorax, canine abdomen, and feline thorax emphasizing their strengths in detecting abnormalities with high sensitivity and accuracy. The designed prompts prioritize patient safety by erring on the side of caution, ensuring that even subtle abnormalities are flagged for further investigation, rather than risking an incorrect classification as normal.

While both models achieved strong overall performance, slight variations across conditions and datasets suggest areas for further improvement. For example, refining prompt designs, incorporating more diverse training data, and optimizing the models for specific conditions could further enhance their reliability and adaptability. Notably, Gemini Flash 1.5 demonstrated a slight edge in performance across two of the three datasets, highlighting the potential for model-specific optimization.

These findings indicate that AI tools can play a very useful supportive role for practitioners and provide a platform on which to base future clinical workflow integration of such models. AI models can contribute to better patient and practitioner outcomes by supporting diagnostic accuracy and reducing the possibility of diagnostic oversight. In practical application, however, ongoing refinement is necessary to decrease false positives and ensure adaptability across various clinical scenarios.

The innovative potential is evident in these models, but it also points toward a success factor: collaboration among AI researchers, veterinarians, and radiologists. Future research, by aligning domain expertise with the advances in AI, could address real-world challenges that will let these tools reach their full potential. As this field evolves, these efforts could revolutionize how veterinary diagnostics are performed-safer, more efficient, and ultimately more effective for all patients.

Appendix A: Condition Definitions

Term	Definition
Bronchiectasis	A condition where airways are damaged, leading to widened, floppy, and scarred tubes.
Bronchitis	Swelling and irritation of the bronchial tube walls.
Cardiomegaly	Abnormal heart enlargement, which may be localized to the left, right, or both sides.
Diseased Lungs	A broad term for any abnormality or condition impacting lung tissue.
Esophagitis	Irritation or swelling of the esophageal lining.
Fe Alveolar	Areas of the lung appear cloudier due to partial or complete alveolar filling.
Focal Caudodorsal Lung	Describes localized changes in the rear-upper lung area, such as masses or infiltrates.
Focal Perihilar	Refers to findings near the lung hilum, such as localized masses or infiltrates.
Hypoplastic Trachea	A congenital condition where the trachea is underdeveloped or narrower than normal.
Interstitial	Pertains to the connective tissue of the lungs; abnormalities cause a specific radiographic pattern.
Left-Sided Cardiomegaly	Enlargement primarily affecting the left chambers of the heart.
Pericardial Effusion	Fluid accumulation in the sac encasing the heart.
Perihilar Infiltrate	Radiographic findings of abnormal opacities near the lung hilum.
Pleural Effusion	Fluid build-up in the space between the lungs and the chest wall.
Pneumonia	Infection-induced inflammation of lung tissue, often caused by bacteria, viruses, or fungi.
Pulmonary Hypo inflation	Reduced lung expansion or areas of collapse visible on imaging.
Pulmonary Nodules	Small, rounded opacities in the lung, which may be benign or malignant.
Pulmonary Vessel Enlargement	Enlargement of pulmonary arteries or veins, visible on imaging.
Right-Sided Cardiomegaly	Enlargement primarily affecting the right heart chambers.
Redundant Tracheal Membrane	A condition where the dorsal tracheal membrane is excessively loose, narrowing the airway.
Thoracic Lymphadenopathy	Swelling of lymph nodes located within the thoracic region.

Ascites	Excess fluid accumulation in the abdominal cavity, causing noticeable swelling.
Colitis	Inflammation of the colon or large intestine.
Gastritis	Irritation or inflammation of the stomach lining.
Hepatomegaly	An increase in liver size beyond normal parameters.
Liver Mass	An abnormal growth or tumor located within or on the liver.
Microhepatia	An unusually small liver compared to normal size.
Pancreatitis	Swelling or inflammation of the pancreas.
Small Intestinal Obstruction	A blockage that restricts the passage of contents through the small intestine.
Splenic Mass	A lump, tumor, or abnormal growth in the spleen.
Splenomegaly	An enlarged spleen, often due to various underlying conditions.

B Appendix: Code Repository Link

https://github.com/SharvenRane/LLM-Diagnostic-Tool-for-Veterniary-Radiology-Text-Data

References

Manuel Duarte Lobo, Artificial Intelligence in Teleradiology, Handbook of Research on Instructional Technologies in Health Education and Allied Disciplines, 10.4018/978-1-6684-7164-7.ch004, (80-104), (2023).

Ana Inês Pereira, Pedro Franco-Gonçalo, Pedro Leite, Alexandrine Ribeiro, Maria Sofia Alves-Pimenta, Bruno Colaço, Cátia Loureiro, Lio Gonçalves, Vítor Filipe, Mário Ginja, Artificial Intelligence in Veterinary Imaging: An Overview, Veterinary Sciences, 10.3390/vetsci10050320, **10**, 5, (320), (2023).

Leah Kathleen Pomerantz, Mauricio Solano, Eric Kalosa-Kenyon, Performance of a commercially available artificial intelligence software for the detection of confirmed pulmonary nodules and masses in canine thoracic radiography, Veterinary Radiology & UltrasoundVeterinary Radiology & UltrasoundVeterinary Radiology & Ultrasound, 10.1111/vru.13287, **64**, 5, (881-889), (2023).

Corrigendum: Accuracy of artificial intelligence software for the detection of confirmed pleural effusion in thoracic radiographs in dogs, Veterinary Radiology & UltrasoundVeterinary Radiology & UltrasoundVeterinary Radiology & Ultrasound, 10.1111/vru.13202, **64**, 1, (155-155), (2022).