Analyzing the Effect of Unsupervised Reinforcement Learning Techniques for Biomedical Condition Classification

Ben Cote
University of Washington
bpc23@uw.edu

Cory Clemmons
Vetology Innocations LLC
Cory.Clemmons@vetology.net

Laxman Shankar Maheshkar
Vetology Innocations LLC
laxman.maheshkar@vetology.net

Sharven Rane
Independent Researcher
sharvenrane7@gmail.com

Abstract

This study evaluates the performance of zero-shot and reinforcement learning-based large language model (LLM) systems on a binary classification task involving radiology reports. Specifically, the models classified a target list of medical conditions as normal or abnormal based on their context within the report. The aim was to assess the utility of LLMs for data science tasks that require domain-specific knowledge, as well as studying the differences between single model and multi-model LLM systems.

Two approaches were compared: (1) a Zero-Shot model using a single prompt, and (2) the Reflexion model, employing an unsupervised reinforcement learning loop with three LLM agents. The hypothesis was that Reflexion's iterative self-analysis would improve classification accuracy. Contrary to this hypothesis, the Zero-Shot model outperformed Reflexion on all three datasets, with an average accuracy difference of 1.6%.

As shown in Table 2, both approaches achieved high accuracy, but the findings highlight challenges in using unsupervised reinforcement learning loops for error correction due to the model's struggle to adhere to task-specific instructions. These results underscore the limitations of unsupervised reinforcement learning for text classification tasks and emphasize the need for careful supervision to minimize deviations from specified objectives. While LLMs show promise for context-sensitive classification, their deployment in medical applications requires thorough oversight to ensure reliability.

Table 1: Results Summary

Dataset	Zero-Shot Acc.	Reflexion Acc.
Canine Thorax	96.7%	94.3%
Canine Abdomen	95%	93.6%
Feline Thorax	97.9%	96.9%
Total	96.5%	94.9%

1 Introduction

This paper describes the development and testing of multi-agent Large Language Model (LLM) systems tasked with classifying medical conditions from radiologist reports. It is an investigation of Agentic Workflow techniques, using multiple LLM agents to handle a task, and evaluating the impact of reinforcement learning on LLM systems. This task can be described as biomedical Named-Entity Recognition (BioNER), the automatic recognition of biomedical entities from natural language text (Luo et al., 2023) combined with text classification to determine whether or not the condition mentioned in the medical report reflects any abnormalities.

Both Named-Entity Recognition and text classification are well-defined objectives for natural language processing, but the method for analysis has changed over time. From context-based classifiers like conditional random fields to the more recent BioBERT, a transformer model trained on labeled medical data (Lee et al., 2020), natural language processing has adapted to the latest advancements in machine learning. Yet, these models have relied on large quantities of manually annotated gold-standard data in order to produce accurate classification predictions. This poses a problem for the biomedical domain because annotation requires skilled workers that have expertise in the specific field. As such, these models have struggled with accuracy and robustness (Luo et al., 2023). The recent developments in deep learning techniques and improved LLMs like GPT-4 (Achiam et al., 2024) present a possible fix for these issues, because they are capable of unsupervised learning without pre-labeled data. While these LLM models have their own difficulties with nuanced comprehension and the use of abbreviations and acronyms (Monajatipoor et al., 2024), careful prompt engi-

neering may address these drawbacks.

Additionally, Agentic Workflow processes are able to improve model output by breaking a task down into smaller component processes and using multiple LLM agents to handle each subtask. The focus of this paper is to analyze the benefit that reinforcement learning can add to an agentic workflow. To do so, a standard zero-shot process is compared to one that uses a reinforcement loop to repeatedly evaluate and correct its errors until the model is satisfied with the classification decisions it makes. Given the complex context-dependent nature of medical condition classification, the anticipated result is that reinforcement learning will have a significant improvement on classification accuracy. If this hypothesis is supported, it will demonstrate how creative use of off-the-shelf LLMs can enable small data science teams to produce high-quality models. For this reason, this short-term project through Vetology investigates how LLM-based multi-agent models can be leveraged to identify abnormal medical conditions from radiologist reports.

Vetology is a San Diego-based company in the medical technology field founded in 2010 with the original purpose of providing a platform for veterinarians and radiologists to share cases and images. However by 2017 they began to investigate ways to integrate machine learning into their platform to create advanced, automated radiograph analysis. Each year the demand for teleradiology increases, but the number of certified radiologists graduating from the American College of Veterinary Radiology cannot keep up with this demand (Cima, 2018). The result is that radiologists have an overwhelming case load, wait times have increased, and the price of services have gone up. This unsustainable growth provides a great opportunity for a company like Vetology to use machine learning and computer vision models to train an automated radiograph analysis system that can provide reports in minutes, giving patients an accurate approximation of a traditional radiology report that may take several days to be completed.

Currently, there is a 92% match rate between the Vetology's virtual radiologist and radiologists' findings, and Vetology is constantly working to improve model performance (Kim et al., 2022). Further, both the American College of Veterinary Radiology and the American College of Radiology have embraced that the responsible development of 'AI applications' may play a role in the future of radiology, especially as demand continues to grow. Vetology has taken its role in ethical AI development very seriously; the leadership team is comprised of radiologists and tech executives who understand both medical and technical fields and seek to join them through their products.

2 Dataset and Data Annotation

There are three datasets in this project, each representing a different animal or body location for the radiograph. The datasets are canine thorax reports, canine abdomen reports, and feline thorax reports, and the conditions in each of these datasets are listed in Appendix A. Each dataset is arranged in the same format, in which each row represents a patient identified by a case ID number, followed by a report written about the patient's radiology scan. Each of these reports is divided into radiologist findings, conclusions, and recommendations for the patient. The canine thorax dataset has 4000 cases, the canine abdomen dataset has 3751 cases, and the feline thorax dataset has 1875 cases. However due to constraints around model runtime, personnel limitations, and financing this project, this project was limited to a subset of 50 reports per dataset, 150 reports total. After a week dedicated to learning the medical terminology relevant for identifying medical conditions in radiograph reports, I began to analyze the data by-hand. All of these reports were independently scored by myself and another data scientist on this project. We separately read through each of the 150 reports and manually produced classification decisions for the list of either twenty-one (for canine and feline thorax) or ten (for canine abdomen) conditions targeted by the model, depending on the dataset. Following our individual data classifications, we then compared our results through a peer review process, identifying and subsequently talking through each disagreement and coming to a conclusion on every decision in order to create one set of agreed-upon gold standard labels. These labels form the standard against which my model results are evaluated.

3 System Overview

Figure 1 shows the flow of our system, from input configuration through data processing and classification, outputting both the modified original dataframe now containing classification decisions for each condition in every report, and a confu-

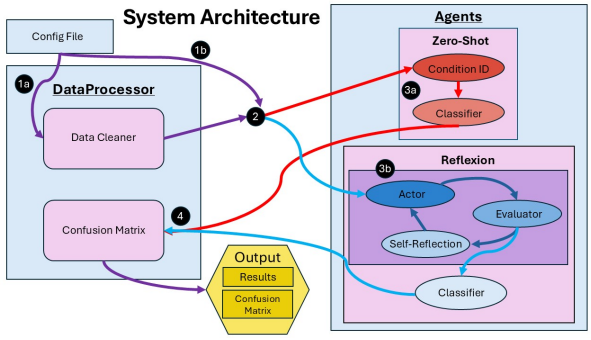

Figure 1: A graphical depiction of the Medical Condition Classification System

sion matrix detailing the the model's performance against the gold standard labels. In this system, the configuration file specifies whether the data must be processed (path 1a) or whether it is preprocessed (1b), includes filepaths for input data and output, the number of reports to analyze, the workflow technique that should be used for the analysis, and which LLM model should be used. The configuration file specifies either a Zero-Shot workflow (3a) or the reinforcement learning Reflexion workflow (3b), and both workflows produce the same output format, which is then added to the original dataframe (4) and an output confusion matrix is generated.

4 Approach

Our approach has two major paths: Zero-Shot and Reflexion. The system begins by performing some minor processing of the dataset, before choosing either the Zero-Shot approach of the Reflexion approach for condition classification. Following this choice, the results are compiled into a confusion matrix for easier visualization.

4.1 Data Cleaner

This stage of the model is designed to ensure that the input data is uniform before it reaches the classification systems. This involves compiling the findings, conclusions, and recommendations columns in the original dataframe into a single 'Report' column. These changes streamline the data input process for calling the large language model. Additionally, this function ensures uniform labels of 'normal' or 'abnormal' for each pre-classified label. Some conditions used the labels of 'enlarged' and 'normal,' but to keep the model classification standardized the cells labeled 'enlarged' were changed to 'abnormal'. It also replaces empty cells with the label 'normal,' fixing an upstream bug in the dataset that only classified conditions when they were abnormal. The reports themselves are not modified at all in data processing. This is because there is already a great variety of report formats, so a standardization process would likely remove data from some reports without providing a clear benefit.

4.2 Zero-Shot Workflow

The Zero-Shot approach–meaning that the LLM performs its task without any prior examples or training–that was implemented in this project is a standard call to an LLM agent using a prompt. In this case, the LLM is provided with a prompt divided into two sections: system information and human information. The system information provides context for the LLM agent about its role and task, while the human information provides the input that the system instructions will be acting upon. The system section of the prompt explains that the LLM is acting as a medical professional tasked with classifying medical conditions as 'normal' or 'abnormal' based on a radiologist report that will be provided later. The system section also includes a list of instructions for classifying conditions, a list of definitions for each condition, and an example output with the desired format. The human section of the prompt contains the actual radiologist report for classification. Both of these prompt sections are provided to the LLM agent, and it produces a list of each specified condition and its classification of normal or abnormal in the provided report.

Individual calls made to LLMs are quite powerful, but they can have limited success when completing complex tasks given that the more instructions presented in a prompt, the greater the chance that a mistake is made. Furthermore, when the task given to the model requires a higher level of reasoning, single LLM calls can struggle regardless of the size of the language model (Rae et al., 2022).

4.3 Reflexion Workflow

To combat the accuracy problems that zero-shot approaches may produces, prompting strategies such as the Chain of Thought approach use multiple LLM agents that each perform an intermediate step in the reasoning process (Wei et al., 2023). This creates more calls to LLM agents, but each has a simpler task, and the output of one call can be used as input for the next call.

The Reflexion workflow implemented in this system is based on a framework of the same name (Shinn et al., 2023). Reflexion is a strategy of unsupervised reinforcement learning, in which the LLM system evaluates its own output and offers feedback to itself as input in order to improve output results over subsequent reinforcement loops. This kind of evaluation process is foundational in machine learning and optimization with algorithms like gradient descent, however the evaluation is usually numerical, and the feedback is a mathematical weight adjustment that helps optimize the algorithm. Reflexion takes a different approach, utilizing large language models' strength at analyzing textual data to provide linguistic feedback that is then re-fed into the original task's prompt. The linguistic aspect of this framework lends itself better to text-based LLM prompts, and allows the feedback to be more specific.

Our Reflexion model contains three separate LLM agents: the Actor, the Evaluator, and the Self-Reflector. The system prompts for each agent describe their individual role on this team of medical professionals, along with providing example output formats to standardize the system's output. The human prompts contain the information necessary to perform their given task.

The Actor is the primary agent in this task. Its job is to read a radiologist's report and generate condition classification decisions. The Actor's prompts include all the information from the Zero-Shot Workflow, with one additional piece allowing space for additional instructions to be added later. This leaves room for the Self-Reflector's feedback to be integrated into the Actor's prompt, therefore reinforcing its classification decisions based on an evaluation of its prior output. The Evaluator agent is tasked with reviewing the radiologist report and the classification decisions made by the Actor agent, and producing a list of condition classifications that it disagrees with, along with an explanation of why it finds the Actor's assessment unsatisfactory. Then, the Self-Reflector is provided with the initial report, the Actor's classifications, the Evaluator's disagreements, and optionally the prior feedback the Self-Reflector produced in the last reinforcement loop. This prior feedback is useful for ensuring that when the Self-Reflector generates its feedback for the Actor, it does not repeat the same instructions or fully contradict its prior feedback in the case of an uncertain diagnosis. The Self-Reflector is also capable of deciding to withhold specific disagreements from the Evaluator's report, if it decides that there is not enough justification. This step, like a peer-review, helps add a second opinion that can challenge any cases where one LLM agent provides an output that does not align with the prompt instructions.

This reinforcement learning loop will end either when the Evaluator finds zero disagreements with the Actor's decisions, or when a set number of loops have occurred. This was limited to five reinforcement loops for the scope of this project. If the maximum number of loops have been reached, the model will return the Actor output that the Evaluator has found with the fewest number of disagreements with. This ensures that the best result of the reinforcement learning loop becomes the output, because it is not guaranteed that the reinforcement learning loop always directly improves the model's results.

4.4 Classifier

The classifier component of the system compares the model's final output classifications against a list of manually scored gold labels for each report. The classifications are as follows: True Negative (Gold Label: normal, Model Label: normal), True Positive (Gold Label: abnormal, Model Label: abnormal), False Negative (Gold Label: abnormal, Model Label: normal), and False Positive (Gold Label: normal, Model Label: abnormal). Each condition in the report receives a classification, and these lists of classifications are then put into a confusion matrix to better visualize which conditions the model struggles with classifying the most.

4.5 Prompt Engineering

A significant portion of this project was dedicated to viewing model results, comparing them to the gold standard datasets, and adjusting the LLM prompts provided to each agent. These adjustments took a variety of forms, from clarifying formatting and task instructions to providing condition-specific instructions to help the agent correctly diagnose a condition, to updating condition definitions to include common phrases used in radiograph reports. In some cases, specific instructions were included to the Evaluator prompt to ensure that it did not second-guess classification decisions made on a condition, because the Evaluator consistently misclassified it. While these prompt tunings are ad-hoc, they are based explicitly on trends observed

Table 2: Results Summary

Dataset	Zero-Shot Accuracy	Reflexion Accuracy
Canine Thorax	96.7%	94.3%
Canine Abdomen	95%	93.6%
Feline Thorax	97.9%	96.9%
Total	96.5%	94.9%

in the model's decisions, and therefore have an important role in providing guard rails to the LLM model.

This process, while highly important, produces inconsistent results. With hundreds of repeated calls to these LLM agents that are simply stochastic parrots (Bender et al., 2021) generating text based on complex probabilities, large language model results are still prone to hallucinations and incorrect outputs. These errors can be reduced through clear and carefully structured prompts, but there is always a possibility that the model won't follow the provided instructions. This variability means that a LLM-based model run several times on the same dataset may provide slightly different outputs, so there is a level of standard error to be expected from model results.

5 Results

The Reflexion and Zero-Shot results display a similar pattern across each of the three datasets. For the canine thorax model, the Zero-Shot method produced an average accuracy score of 96.7%, while the respective Reflexion method produced an accuracy score of 94.3%. For the canine abdomen model, the Zero-Shot method had an average accuracy of 95% compared to the Reflexion accuracy score of 93.6%. For the feline thorax model, the Zero-Shot method produced an accuracy of 97.9%, and the Reflexion method had an accuracy of 96.9%. In all of these cases, the Zero-Shot method outperformed the Reflexion workflow by a small margin, with the largest difference in model accuracies being in the canine thorax dataset with a difference of 2.4%. The overall average accuracy difference between models was 1.6%. Tables 3, 4, and 5 display the rates of true positive (TP) cases for correctly classified abnormalities, false negative (FN) for normal cases incorrectly classified as abnormal, true negative (TN) cases for correctly classified normal conditions, false positive (FP) for abnormal cases incorrectly classified as normal, and an accuracy score representing the percentage of cases that were classified correctly regardless of abnormality. The bolded accuracy numbers indicate the model that had a better performance on the given condition, and when both methodologies produced the same classification accuracy, neither workflow's accuracy is in bold.

While not all of the results are included in this paper, these two methodologies were tested using three sets of LLM models: GPT-4-mini, Llama 3, and Claude. The results shown are from the OpenAI GPT-4-mini model, however the performance differences between different LLMs were insignificant in this particular classification task. The implications of this will be discussed further in the Discussion section below.

In addition to a higher accuracy, the lack of a reinforcement loop within the Zero-Shot method means that it generates its results with fewer calls to the LLM agents which means a lower unit cost for the model, and can classify the same number of reports in under half of the amount of time that it takes the Reflexion workflow.

6 Discussion

Across all datasets, the Zero-Shot workflow outperformed the unsupervised reinforcement learning of the Reflexion model. This was unexpected, given that the Reflexion model outperformed the Zero-Shot method upon initial implementation. Likely a result of the Reflexion workflow's opportunity for disagreement identification and revision of classification decisions, it was able to catch some initial errors and correct them, thus producing higher results than the Zero-Shot model. However through the repeated process of analyzing model results on the training dataset, the prompts provided to these LLM agents were adjusted to address repeated errors and misunderstandings. For example, the model tended not to classify cardiomegaly as abnormal when it had qualified left-sided or right-sided cardiomegaly as abnormal. Since cardiomegaly is a more general term, it should also be classified as abnormal when one of the more specific classifica-

Table 3: Condition Classification for Canine Thorax Reports

Condition	Zero-Shot					Reflexion				
	TP	FN	TN	FP	Accuracy	TP	FN	TN	FP	Accuracy
bronchiectasis	0	0	50	0	**100%**	0	0	49	1	98%
bronchitis	16	1	32	1	**96%**	11	6	30	3	82%
cardiomegaly	9	0	40	1	**98%**	5	4	41	0	92%
diseased lungs	25	0	21	4	**92%**	23	2	21	4	88%
esophagitis	1	0	49	0	**100%**	1	0	49	0	100%
focal caudodorsal lung	6	0	35	9	82%	2	4	43	1	**90%**
focal perihilar	3	0	46	1	**98%**	0	3	47	0	94%
hypoplastic trachea	0	1	49	0	98%	0	1	49	0	98%
interstitial	11	0	36	3	**94%**	7	4	33	6	80%
left-sided cardiomegaly	0	6	44	0	88%	0	6	44	0	88%
pericardial effusion	1	0	49	0	**100%**	0	1	49	0	98%
perihilar infiltrate	4	0	45	1	98%	3	1	46	0	98%
pleural effusion	2	0	48	0	**100%**	0	2	48	0	96%
pneumonia	4	0	44	2	96%	3	1	46	0	**98%**
pulmonary hypoinflation	4	0	45	1	98%	4	0	46	0	**100%**
pulmonary nodules	6	0	44	0	**100%**	3	3	44	0	94%
pulmonary vessel enlargement	1	0	49	0	100%	0	1	49	0	98%
right-sided cardiomegaly	0	2	48	0	96%	0	2	48	0	96%
rtm	2	0	47	1	**98%**	1	1	47	1	96%
thoracic lymphadenopathy	3	0	47	0	**100%**	2	1	47	0	98%
VHS v2	0	0	50	0	100%	0	0	50	0	100%

Table 4: Condition Classification for Canine Abdomen Reports

Condition	Zero-Shot					Reflexion				
	TP	FN	TN	FP	Accuracy	TP	FN	TN	FP	Accuracy
ascites	3	0	47	0	**100%**	3	0	46	1	98%
colitis	12	0	32	6	**88%**	9	3	34	4	86%
gastritis	8	0	38	4	**92%**	4	4	41	1	90%
hepatomegaly	6	0	42	2	96%	5	1	44	0	**98%**
liver mass	0	1	46	3	**92%**	0	1	45	0	98%
microhepatia	4	1	45	0	98%	4	1	45	0	98%
pancreatitis	11	1	35	3	**92%**	11	1	33	5	88%
small intestinal obstruction	6	0	43	1	**98%**	5	1	42	2	94%
splenic mass	3	1	46	0	98%	3	1	46	0	98%
splenomegaly	1	0	47	2	96%	0	1	48	1	96%

Table 5: Condition Classification for Feline Thorax Reports

Condition	Zero-Shot					Reflexion				
	TP	FN	TN	FP	Accuracy	TP	FN	TN	FP	Accuracy
bronchiectasis	0	0	50	0	100%	0	0	50	0	100%
bronchitis	22	0	27	1	**98%**	12	10	28	0	80%
cardiomegaly	12	0	38	0	**100%**	9	3	38	0	94%
diseased lungs	28	4	17	1	**90%**	27	5	17	1	88%
esophagitis	2	0	48	0	**100%**	1	1	48	0	98%
Fe Alveolar	0	0	50	0	100%	0	0	50	0	100%
focal caudodorsal lung	4	0	43	3	**94%**	2	2	44	2	92
focal perihilar	0	0	49	1	98%	0	0	50	0	**100%**
hypoplastic trachea	0	0	50	0	100%	0	0	50	0	100%
interstitial	4	0	44	2	96%	3	1	46	0	**98%**
left-sided cardiomegaly	0	1	49	0	98%	0	1	49	0	98%
pericardial effusion	0	0	50	0	100%	0	0	50	0	100%
perihilar infiltrate	4	0	45	1	98%	3	1	46	0	98%
pleural effusion	0	0	49	1	98%	1	0	49	0	**96%**
pneumonia	1	0	47	2	96%	0	1	48	1	96%
pulmonary hypoinflation	1	0	49	0	100%	1	0	49	0	100%
pulmonary nodules	6	1	42	1	96%	6	1	43	0	**98%**
pulmonary vessel enlargement	2	0	48	0	100%	2	0	48	0	100%
right-sided cardiomegaly	0	1	49	0	98%	0	1	49	0	98%
rtm	0	0	50	0	100%	0	0	50	0	100%
thoracic lymphadenopathy	1	0	47	2	96%	0	1	48	1	96%

tions is also abnormal. However, the presence of a generalized cardial enlargement should result in a classification of cardiomegaly as abnormal, but classify left-sided and right-sided cardiomegaly as normal. To address this concern, the following special instruction was added to the Zero-Shot and Reflexion Actor prompts: "For cardiomegaly: If left-sided or right-sided cardiomegaly is classified as 'abnormal,' classify cardiomegaly as 'abnormal' as well. If cardiomegaly is classified as 'abnormal' without specifying left or right, classify both left-sided and right-sided cardiomegaly as 'normal.'" Another general prompt modification that helped improve the results of both workflows was the inclusion of common key words in the radiologist reports that indicated abnormality: "Examples of terms that should typically lead to an 'abnormal' classification include 'inflammation,' 'enlargement,' 'opacity,' 'thickening,' 'mass,' or any descriptors indicating deviation." Through the inclusion of these terms, the models were better able to predict abnormalities. The addition of prompt modifications like those above were the only enhancements provided to the Zero-Shot model to enhance performance. The process of adjusting the Reflexion workflow involved more steps given that there were three agents working together. Analyzing common mistakes and patterns in how the Reflexion model generated results helps to identify both the strengths and weaknesses of LLM-based unsupervised reinforcement learning.

6.1 Dataset Size Limitations

It is important to note that given the small dataset size of 50 cases per dataset, this paper's results may not reflect the same trends as a much larger analysis. Therefore, the conclusions of this paper may not be generalizable beyond its original scope. However besides the prompts to guide classification, most of the decision making comes from the pretraining of the LLM in use. These LLMs have been trained on massive datasets, with GPT-3 having a training set of 499 billion tokens (Brown et al., 2020), so the model's decisions should be consistent across both small and large datasets. Regardless, this project is meant to be a exploratory comparison of two LLM workflows, and both workflows were tested on the same dataset, so there is still validity in assessing their differences.

6.2 Minimal Model Differences

The fact that results did not differ significantly across the GPT-4-mini, Llama 3, and Claude models suggests that despite their differences, each model has similar capabilities on a text classification task. Even multi-model approaches such as using GPT-4-mini as the Actor and Self-Reflector and Claude as the Evaluator did not produce different model results. On occasion, the wording of a particular prompt would need to be adjusted for a different LLM; Claude's model outputs did not match the format stated in the prompt, so it required additional instructions to ensure the desired output was achieved. There may be use cases where one LLM outperforms others, but this project did not explore any such tasks.

6.3 Decision Convergence Issues

In gradient descent-based reinforcement learning, parameter weights are adjusted repeatedly to help the model approach a more optimal result. When changes are too large, it can result in overshooting

Table 6: Evaluator Disagreement Log for Bronchitis

Loop	Agent Classification	Evaluator Disagreement
1	Abnormal	The report only mentions a mild generalized bronchointerstitial pulmonary pattern, which does not provide clear evidence of bronchitis. Bronchitis should be classified as normal.
2	Normal	The report identifies a mild diffuse bronchointerstitial pulmonary pattern, which suggests the presence of bronchitis. Bronchitis should be classified as abnormal.
3	Abnormal	The report only describes a mild generalized bronchointerstitial pulmonary pattern, which does not provide clear evidence of bronchitis. The report states this pattern may indicate various conditions, but does not definitively diagnose bronchitis. Bronchitis should be classified as normal.
4	Normal	The report identifies a mild diffuse bronchointerstitial pulmonary pattern, which suggests the presence of bronchitis. Bronchitis should be classified as abnormal.

a local inflection point. When this happens repeatedly, the model oscillates over the inflection point without ever converging. There are ways to handle this in gradient descent by making the learning rate smaller to ensure that overstepping cannot occur. However, in Reflexion's text-based reinforcement learning there are no weights to adjust. Instead, adjusting the learning has to take place through prompt adjustments specifying what to mention or avoid mentioning in your output and example outputs to provide guidance. While this approach is more flexible and tailored to LLMs, it lacks the specificity and control that makes most reinforcement learning techniques reliable.

One example of this lack of control can be observed through a habit of oscillation in the reinforcement learning loop. When the radiologist report is ambiguous about whether a condition is abnormal or not, the model struggles to converge on one decision. Using output logs to capture the Evaluator agent's disagreements with the Actor's decisions reveals an unending cycle switching between normal and abnormal classifications of the condition bronchitis in a canine thorax.

Table 6 demonstrates how the reinforcement loop does not help the model to converge on one answer. Instead, the Evaluator disagrees with the Actor's decision no matter what, because it can produce evidence that contradicts both a 'normal' and 'abnormal' classification. It is possible that this pattern could be avoided using more advanced memory systems so that the model does not contradict its previous disagreement, but that would reduce the utility of reinforcement learning. Overall, the unsupervised reinforcement learning strategies used in this project are still powerful techniques that catch errors and provide valid disagreements, but they are not a substitute for supervised or semi-supervised reinforcement learning methods that use non-LLM feedback to guide the learning process. I anticipate that a semi-supervised system would be able to outperform both the Zero-Shot and Reflexion models.

6.4 Prompting Limitations

A great strength of large language models, their ability to generate highly probable outputs based on complex prediction strategies, is also one of their largest problems. Over the course of this project, the prompts for LLM agents were revised continuously in order to standardize their output format, provide specific instructions to curb repeated errors, and to clarify the agent's task. Together, these revisions resulted in a more standardized model output and improved the classification accuracy of the model. Yet I also ran into limitations with how much my prompts could effect the model's internal text generation strategies. For instance, my Actor prompt instructions state that the agent output should "[i]nclude all 21 conditions in your response, in alphabetical order as provided." and then provides a list of all twenty one conditions with definitions. It is explicitly states that the agent's task is only to classify the twenty one conditions, and not to classify any additional conditions. Despite this, the LLM output for one feline thorax case contained eleven additional conditions that it had classified: c2-3 intervertebral disc disease, thoracic/abdominal imaging, atlantoaxial joint subluxation, dens, intervertebral foramina, and six other conditions. This is in clear defiance of the prompt instructions. It did not end up having an adverse effect on the model because I included an additional checkpoint in my code that removed any conditions not explicitly requested, but this serves to show that LLM outputs cannot be controlled, even with the most careful prompting and low temperature to maximize consistency.

The addition of extra conditions may not prove a serious threat to the model's functioning, but in my model results there are also examples of the LLM agent's outputs violating explicit task instructions. This project is built to provide medical condition classifications, therefore it is important that the classifications are overly-cautious because it is safer to generate false positives that can be corrected with a second opinion than to generate false negatives that fail to flag a potential problem. This was a key consideration in the generation of gold standard datasets and LLM prompts, where a condition should be marked as abnormal if there is a possibility that there is an abnormality, regardless of how unlikely it is. This posed a conflict for the reinforcement learning loop, because the Actor would generate decisions based on these strict guidelines and then the Evaluator would disagree despite being provided the same strict guidelines. The conflict between my provided instructions and the task to identify disagreements was something the model was unable to overcome. This can be seen in the following Evaluator justification for disagreeing with the Actor's decision on colitis: "the

teammate has classified colitis as abnormal, but the report mentions non-specific gastrointestinal tract appearance such as from enteritis or colitis, suggesting that colitis should be classified as normal since it is not definitively indicated as present." The Evaluator should not have disagreed with the Actor's decision, because its task instructions include a special instruction for classifying colitis stating that "[i]f the report mentions non-specific findings that indicate colitis, classify colitis as abnormal."

One reason that the Zero-Shot model outperformed the Reflexion model is that it did not have a system in place to double-check model decisions. In this particular case, Reflexion's use of a second LLM opinion hurt model performance because it disobeyed instructions to mark potential abnormalities as 'abnormal' even without irrefutable evidence. In other domains and tasks, the Reflexion workflow's reinforcement loop could serve as a tool for checking for errors, identifying and correcting hallucinations, or providing a moderating effect on an agent's extreme decisions. But this suggests that while the transformer model's internal 'reasoning' can be guided by specific prompting, it cannot be overridden. Therefore unless a specific task's goals are aligned with the model's reasoning, there will be a logical dissonance represented in the results, and the reinforcement learning's moderating effect may either benefit or harm the model's results, depending on the task's goals and reasoning.

7 Conclusion

Overall, this project demonstrates that large language models are capable of generating highly accurate results on text classification tasks, even in niche domains such as veterinary radiology. Their predictive power is useful for identifying common language patterns that indicate when conditions are normal or abnormal, and assigning the correct decision to those conditions. However, they are unable to follow prompt instructions that require the LLM to contradict their internal 'reasoning' to produce more cautious results. Models tend to follow the prompt instructions clearly, but the errors generated during this project are results of the predictive systems underlying large language models. LLMs are not capable of reasoning, and therefore these mistakes are inevitable. Perhaps they are infrequent enough that they don't have a large effect on model performance over thousands or millions of classification decisions, but these large transformer models should not be treated as intelligent just because their output looks reasonable. While unsupervised reinforcement learning using LLMs reveals their internal biases, these could certainly be corrected using a supervised or semi-supervised reinforcement learning method.

When it comes to LLM approaches for tasks requiring high level reasoning, agentic approaches with multiple LLM agents tend to outperform zero-shot methods (Rae et al., 2022). Yet in this specific task, the Zero-Shot approach performed better than the agentic Reflexion approach because it placed more emphasis on the human-generated prompt and did not have several LLM agents capable of contradicting and overriding each other's decisions. What often moderates model outputs served to harm model results in this case. Like all other techniques in data science, there are benefits and drawbacks to any approach. The difficulty with large language models is that their drawbacks are often hard to identify because neural network's decision layers are hidden within the model, and their outputs are designed to appear correct even when they are not. Moreover, these models are designed to produce slightly different results upon subsequent calls using the same prompt. Even with the lowest temperature settings, LLM-based results will not be as consistent across trials as traditional machine learning models. LLMs should be used for text classification because of their success across domains, their incredible ability to recognize data patterns, and the fact that LLMs outperform non-neural approaches on many tasks, but it places a great responsibility on the programmer to determine the safety and ethics of their use.

Prompt engineering is a never-ending task; there are always small modifications that could improve model results, and it is impossible to pinpoint what prompt adjustments will have a positive model effect because the inner-workings of the LLM transformer systems cannot be observed. This is to say that someone may be able to replicate this exact same study with slightly different results based on different prompts. This exploratory project is just one attempt to understand alternative systems using language models.

A Appendix: Condition Definitions

Canine and Feline Thorax Terminology

Term	Definition
Bronchiectasis	Occurs when the tubes that carry air in and out of your lungs get damaged, causing them to widen and become loose and scarred.
Bronchitis	Inflammation of the bronchial walls.
Cardiomegaly	Enlargement of the heart. Cardiomegaly can be left-sided, right-sided, or generalized.
Diseased Lungs	A general term describing abnormal lung tissue. This catch-all term refers to any condition affecting the lungs.
Esophagitis	Inflammation of the esophagus.
Alveolar (Feline Only)	The existence of broad portions of the lung looking more opaque than normal due to partial or complete alveolar filling.
Focal Caudodorsal Lung	Refers to a specific area in the caudal (rear) and dorsal (upper) parts of the lungs, often used to describe localized findings, such as infiltrates or masses, within this lung region.
Focal Perihilar	A specific area around the lung hilum (where bronchi, blood vessels, and nerves enter the lung), often used to describe localized findings such as masses or infiltrates in this region.
Hypoplastic Trachea	A congenitally underdeveloped or narrowed trachea.
Interstitial	Refers to the lung's interstitial space, which includes connective tissue surrounding alveoli, blood vessels, and airways. When this space becomes thickened or inflamed, it can create an interstitial pattern or a bronchointerstitial pattern on radiographs.
Left-Sided Cardiomegaly	Enlargement of the left side of the heart.
Pericardial Effusion	Accumulation of fluid within the pericardial sac surrounding the heart.
Perihilar Infiltrate	Abnormal radiographic opacities or infiltrates near the lung hilum.
Pleural Effusion	Accumulation of fluid within the pleural space between the lungs and the chest wall.
Pneumonia	Inflammation of the lung parenchyma, typically due to bacterial, viral, or fungal infections.
Pulmonary Hypoinflation	Refers to underinflation or reduced expansion of the lungs, where portions of the lung parenchyma appear collapsed or minimally aerated on radiographs.
Pulmonary Nodules	Small, rounded opacities within the lung, which may be benign (e.g., granulomas) or malignant (e.g., metastases or primary lung tumors).
Pulmonary Vessel Enlargement	Abnormal dilation of the pulmonary arteries or veins.
Right-Sided Cardiomegaly	Enlargement of the right side of the heart.
Redundant Tracheal Membrane	A condition where the dorsal tracheal membrane becomes redundant or loose, often causing a narrowed tracheal lumen.
Thoracic Lymphadenopathy	Enlargement of lymph nodes within the thoracic cavity.
Vertebral Heart Score (Canine Only)	An updated method to assess heart size on radiographs, typically used to identify cardiomegaly, by comparing heart size to vertebral length on lateral views.

Canine Abdomen Terminology

Term	Definition
Ascites	When fluid builds up in the abdomen, or belly, causing swelling.
Colitis	Inflammation of the colon or large intestine.
Gastritis	Inflammation or irritation of the stomach lining.
Hepatomegaly	An enlarged liver, or a liver that's larger than its normal size.
Liver Mass	Abnormal growth of liver cells on or in the liver.
Microhepatia	An abnormally small liver.
Pancreatitis	Inflammation of the pancreas.
Small Intestinal Obstruction	A blockage in the small intestine.
Splenic Mass	A lump or tumor in the spleen.
Splenomegaly	A condition where the spleen is enlarged in size or weight.

B Appendix: Code Repository

This project's repository can be found here on GitHub or directly via this URL: `https://github.com/bpcot23/Vetology_Project`. Prompts and data are not included in this repository due to privacy concerns with Vetology's data.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et al. 2024. Gpt-4 technical report. Technical report, OpenAI.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Greg Cima. 2018. Specialists in short supply. In *JAVMA News, October 15, 2018 Issue*. American Veterinary Medical Association.

Eunbee Kim, Anthony J. Fischetti, Pratheev Sreetharan, Joel G. Weltman, and Philip R. Fox. 2022. Comparison of artificial intelligence to the veterinary radiologist's diagnosis of canine cardiogenic pulmonary edema. *Veterinary Radiology & Ultrasound*, 63(3):292–297.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. 2023. AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, 39(5):btad310.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. Llms in biomedicine: A study on clinical named entity recognition.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby

Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis insights from training gopher.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.