ANALYSING ENVIRONMENTAL EFFICIENCY IN AI FOR X-RAY DIAGNOSIS

Liam Kearns
AuraQ
Malvern
Worcestershire
UK
liamkearnsy@gmail.com

ABSTRACT

The integration of AI tools into medical applications has aimed to improve the efficiency of diagnosis. The emergence of large language models (LLMs), such as ChatGPT and Claude, has expanded this integration even further. Because of LLM versatility and ease of use through APIs, these larger models are often utilised even though smaller, custom models can be used instead. In this paper, LLMs and small discriminative models are integrated into a Mendix application to detect Covid-19 in chest X-rays. These discriminative models are also used to provide knowledge bases for LLMs to improve accuracy. This provides a benchmark study of 14 different model configurations for comparison of accuracy and environmental impact. The findings indicated that while smaller models reduced the carbon footprint of the application, the output was biased towards a positive diagnosis and the output probabilities were lacking confidence. Meanwhile, restricting LLMs to only give probabilistic output caused poor performance in both accuracy and carbon footprint, demonstrating the risk of using LLMs as a universal AI solution. While using the smaller LLM GPT-4.1-Nano reduced the carbon footprint by 94.2% compared to the larger models, this was still disproportionate to the discriminative models; the most efficient solution was the Covid-Net model. Although it had a larger carbon footprint than other small models, its carbon footprint was 99.9% less than when using GPT-4.5-Preview, whilst achieving an accuracy of 95.5%, the highest of all models examined. This paper contributes to knowledge by comparing generative and discriminative models in Covid-19 detection as well as highlighting the environmental risk of using generative tools for classification tasks.

***K*eywords** Covid-19 · machine learning · carbon footprint · green AI

1 Introduction

In recent years, there has been an increase in the commercialisation of machine learning models, especially large language models (LLMs). This has led to a rapid advancement in the capabilities and use cases of artificial intelligence (AI). As these models have grown in their capabilities, their size have increased, requiring a greater level of computational resources. With GPT-4 consisting of more than 1.7 trillion parameters [1], the size of these models can be in the terabyte range. Models of this size are not feasible for individuals or companies with limited resources to host, thus creating a reliance on third parties to host such models.

The medical field has been significantly influenced by advances in AI. Commercialised LLMs, such as ChatGPT and Claude, have been proposed as tools to assist with medical diagnoses [2, 3, 4]. Nevertheless, an overreliance on third-party models is not a universal solution. Sharing data with external providers can cause security risks and infringe on regulations if malpractice risks patient harm [5, 6]. LLMs can also falter with simple, case-specific tasks despite showing high confidence in the outputs given [7]. Furthermore, using API calls to an external provider comes with disadvantages, such as higher latency and costs, which can increase the carbon footprint of medical applications. This is a prominent issue in the medical field, with medical imaging emitting high levels of greenhouse gases [8]. As AI

assistance in this field increases, concerns regarding carbon emissions caused by using third-party LLMs will become a prominent issue.

An alternative to third-party models is to deploy models in the environment of an application. Although this reduces latency and gives greater control over data security, resource consumption becomes a major concern and limitation. It is impractical to deploy commercial-sized LLMs in application environments; therefore, local models must be designed and developed to consume less memory. However, reducing model sizes typically means narrowing the scope and accuracy. This trade-off is not suitable for the medical field, as inaccurate disease detection can result in delayed diagnosis and treatment [9]. Consequently, reducing model memory to reduce carbon footprints should not come at the cost of model accuracy.

This paper investigates the performance of LLMs with smaller discriminative models in a classification task. The accuracy and carbon footprints of models are compared to determine whether LLMs are a sustainable tool for specific classification tasks, or whether discriminative models improve environmental sustainability of AI-assisted applications without negatively impacting model accuracy. In the context of this paper, small models and local models are used as interchangeable terms. These terms refer to the discriminative models deployed in the application environment discussed in Section 3.

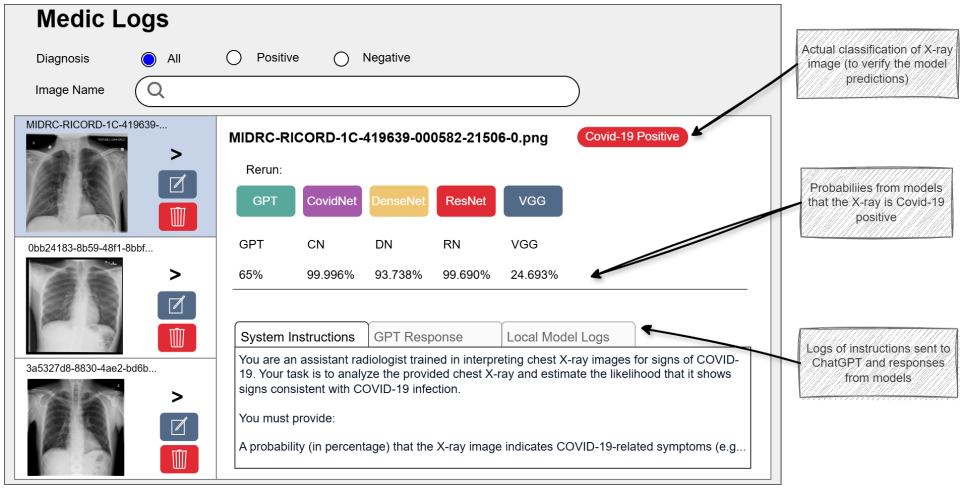


Figure 1: Application using machine learning models to diagnose Covid-19 in chest X-rays.

The scope of this paper is focused on using machine learning models in a medical context for the detection of Covid-19 from chest X-rays. As illustrated in Figure 1, the application used in this paper uses both externally hosted LLMs and local models to determine the probability that an X-ray shows signs of Covid-19. Given the implementation of discriminative models as local models, a classification use case was required to train these models. Although generative models have a larger scope of capabilities than just outputting probabilities, the focus on a single disease facilitates straightforward comparison between models, with the model outputs determining whether an X-ray shows signs of the presence of Covid-19. Because of the specific scope, the results of this paper may not extrapolate to the detection of other diseases, primarily due to data available to train models on other diseases. This scope aims to determine whether local models can outperform LLMs in accuracy and environmental efficiency when given a specific field to work in.

The structure of the paper is as follows. Section 2 highlights state-of-the-art models for X-ray analysis when using LLMs and smaller discriminative models as well as their environmental impact. Next, Section 3 outlines how LLMs and local models will be integrated into a medical application to compare their performance in detecting Covid-19 from X-ray images. Section 4 discusses these results, evaluating their suitability to be used in medical diagnosis before identifying limitations within this study.

1.1 Contributions

In this paper, an analysis of model accuracy and environmental impact is focused on when integrated into medical applications. This outlines a benchmark study of existing model infrastructure and its use in the scope of Covid-19 detection. The major contributions of this paper are as followed:

- Demonstrating major accuracy and carbon footprint concerns when using LLMs to request probabilistic outputs to classify diseases present in X-rays.

- Highlighting both diagnosis accuracy and carbon footprint can be reduced by using local models deployed alongside applications.
- Showing that focusing on reducing carbon footprint over model accuracy can diminish the confidence and accuracy of outputs from smaller models.
- Identifying that knowledge bases for LLMs can increase detection accuracy, but has a varied impact on carbon footprint.

2 Related works

2.1 Large language models

The medical field has recently been introduced to the capabilities of LLMs. Implemented models have shown an accuracy of more than 70% when verified by medical professionals [10]. Notably, ChatGPT has shown accurate identification across various diseases [11]. This underscores the versatility of LLMs, in which models can be used to detect and discuss multiple pathologies.

Although LLMs provide robustness through generalisation, there are risks on the integrity of their outputs. Censorship or bias can occur unintentionally when models are trained on datasets containing censored information [12]. More malicious threats to the integrity of LLM outputs can be caused by prompt manipulation. Although implementing semantic censorship can mitigate malicious prompts, it is important to note that such prompts can be split up and re-worded to disguise malicious intent [13]. Prompt injection is also a vulnerability for some LLMs, where malicious users can make predefined prompts redundant or leak prompts that are hidden from end users [14]. In medical contexts, bypassing LLM safety mechanisms can be dangerous to the health of patients [6]. These vulnerabilities pose a security risk that can be exploited if inputs from end users are directly used in the prompt that is subsequently sent to LLMs. Therefore, it is important to not give unquestioned trust in LLM outputs due to the risks surrounding their training and inference.

Training LLMs for medical purposes is a challenge due to ethical concerns. Medical information is highly sensitive, so mismanagement of data permission can result in serious legal action. To ensure compliance, the algorithms and training procedures of LLMs must be independently reviewed [5]. With commercialised LLM algorithms often not open source, peer-reviewing their compliance with using sensitive patient information is not always possible. This results in less data available for LLMs to train on. Consequently, models are able to detect between healthy and unhealthy medical images, but can struggle with the precise classification of specific abnormalities [15]. Additionally, diagnosing less common conditions results in low accuracy [11]. Without a secure way to train on medical data without compromising patient safety, obtaining accurate medical results from third-party models will continue to be a challenge.

An alternative to fine-tuning LLMs is to use a knowledge base to provide the model with additional information about the medical question being asked. Retrieval augmented generation (RAG) has been used to reduce hallucination in generated medical reports, providing more consistent outputs [16]. RAGs using cosine similarity to retrieve the most relevant information have seen medical analysis with ChatGPT models increase more than 20% [17]. Knowledge bases can also be used to retrieve information from trusted sources, resulting in more reliable advice from medical images [18]. This shows an alternative approach to providing additional information to LLMs without having to go through the process of fine-tuning the model.

Attempting to automate clinical actions through LLMs poses safety concerns. Although existing studies show promising results in high accuracy of LLMs in X-ray analysis, disruption and damage can occur when LLMs output biased or factually incorrect information [3]. LLM performance can improve when the context of a positive diagnosis and recommended next steps are asked within the prompt, but models can overlook important information due to their attention mechanism [2]. Given the impracticability of expecting consistent and accurate outputs from a model, implementing safeguards to prevent unfiltered communication between the model and patients is important to ensure data security and patient safety.

LLMs have the ability to provide explanations for their answers, which can be beneficial when defending clinical decisions. However, the explanations provided by LLMs are not guaranteed to be factual. Reasoning models have struggled to provide accurate outputs in scenarios where the output format is expected to be consistent [7]. In medical scopes, ChatGPT has shown confidence in the reasoning behind its results in relation to X-ray analysis, even in instances where its diagnosis is incorrect [19]. Although other models such as Claude-3.5 Sonnet have shown more consistent results in medical tasks, low sensitivity values highlight that confusion is still apparent in these models [4]. Improvements can be made to minimise LLM hallucinations by verifying quantitative measurements provided in the model outputs [20]. However, these measures rectify errors generated by LLMs rather than fine-tuning LLMs to avoid

these hallucinations from occurring. So, while steps can be taken to rectify fictitious outputs, the risks of hallucinations in model output still pose a threat to medical diagnoses.

2.2 Small models

Although LLMs have seen an increase in focus and improved accuracy, smaller models have not been made obsolete. Notably, LLMs struggle with classification tasks, especially when the correct label is not provided in the input [21]. So, while the versatility of LLMs allows classification tasks to be performed, comprehension of the task cannot be guaranteed. Moreover, in situations where resources are limited and the project has a narrow scope, the use of small models has been shown to be advantageous [22]. In scenarios where training is restricted, smaller models have been shown to outperform larger models [23]. This enhanced performance with reduced resources leads to increased efficiency in training and acquiring outputs from smaller models. In medical scopes, this provides a benefit where substantial data on a novel disease may not be readily available or when there is a lack of funding to train models on energy-intensive equipment.

Neural networks have been shown to provide accurate classifications in medical fields. The VGG model has achieved an accuracy of more than 92% in detecting pneumonia on chest X-rays [24]. These high accuracies are partially obtained by restricting models through classification, whereby outcomes are grouped. Discriminative models of this nature require labelled data for training; however, they have been shown to outperform generative models such as LLMs [25]. Expanding from binary classification to multi-classification (e.g., classifying X-rays as normal, Covid-19, or pneumonia) can reduce accuracy to 82% [26]. So, while neural networks can achieve high accuracies through classification, this also narrows their scope, making their use case specialised.

Covid-Net is a neural network design that focuses on detecting Covid-19 from chest X-rays. The initial version of the model achieved an accuracy of 93.3% [27]. Over time, the open source initiative has increased, resulting in the Covidx CXR-3 dataset, allowing Covid-Net models to reach accuracies of over 96% [28]. Existing model architecture has also been used to accurately detect Covid-19 in X-ray images. A fine-tuned DenseNet model achieved 92% accuracy in detecting Covid-19 despite its reduced complexity compared to other models [29]. This ability to fine-tune models for medical scenarios highlights the robustness of smaller models. Although not as accurate as models like Covid-Net, which are built specifically for Covid-19 detection, the reduced complexity results in less memory usage, making it suitable for resource-constrained environments.

2.3 Environmental efficiency of models

Hosting machine learning models on third-party cloud computing servers can help reduce their carbon footprint through shared resources. With large cloud providers that allow machine learning models to be hosted in low carbon countries in efficient data centres, energy usage can be reduced to have a more positive environmental impact [30]. Moreover, the storage of models on the cloud allows collaboration on a model, reducing the redundancy of creating and training additional models [8]. This shows that actions can be taken when models are deployed in the cloud, improving environmental sustainability.

Smaller discriminative models can be made more environmentally sustainable. Measures to reduce the carbon footprint of discriminative models have included using smaller models during times of high carbon intensity from power sources [31]. However, this means accepting a lower accuracy when the carbon intensity is high. In medical contexts, more environmentally conscious models have been argued to be more competitive due to their reduced training and inference times [32]. However, these models also trade accuracy for a more environmentally conscious approach. This highlights that considerations must be made when focusing on the carbon footprint of discriminative models, whereby accuracy is often negatively impacted to ensure faster model inference.

An increase in input parameters for a model does not necessarily relate to a larger carbon footprint. Factors such as power efficiency in the data centres that store the models and training time mean that models with more input parameters can have a significantly lower carbon footprint [33]. However, language processing models risk being overparameterised, resulting in less efficient training [34]. Therefore, while an LLM with a greater number of input parameters does not necessarily result in an increased carbon footprint, the complexity of these models requires more computationally intensive training compared to other models.

Calculating precise carbon footprints for machine learning is a challenge due to multiple distributed factors. With machine learning models often trained and deployed on external servers, obtaining exact emission data is difficult [33]. Although newer hardware may claim to be more energy efficient, uncertainty in the manufacturing process means that a comprehensive assessment of the full environmental impact remains a challenge [35]. Although precise calculations are difficult, it is evident that interacting with multipurpose models, such as ChatGPT, consumes more

energy than smaller task-specific models [36]. Moreover, LLMs that require greater GPU usage result in increased carbon emissions due to more energy being consumed [37] So, while calculating the precise carbon footprint of models is difficult due to potential unknowns surrounding training and deployment, carbon footprints from model inference are more obtainable.

3 Methodology

Machine learning provides an important tool in Covid-19 detection to improve diagnosis speed and accuracy. However, medical departments are not guaranteed an abundance of computer resources. In addition, there is an increasing concern about the carbon footprint when integrating AI models into medical applications; although LLMs can be inferred through APIs to address limitations of local compute constraints, the use of such large models for a binary classification task may be disproportionate to the resulting environmental impact. This section outlines an application used for Covid-19 detection, in which multiple different models are tested to determine optimal accuracy and energy efficiency.

To compare the environmental efficiency of AI for X-ray diagnosis, both third-party generative and local discriminative models are used. OpenAI and Claude-3.5 models have been applied in previous studies to provide X-ray analysis [11, 4]. LLMs are hosted by various providers and have varying infrastructure, which affects their power consumption [37]. Therefore, to provide a benchmark study, models of different sizes and hardware requirements are analysed.

For discriminative models, Covid-Net is a model built specifically for Covid-19 detection, which makes it a suitable model to compare in this paper [27, 28]. The neural networks DenseNet, ResNet and VGG have shown success in X-ray tasks [24, 29]. These models are smaller in size than Covid-Net. Therefore, testing with these models alongside Covid-Net will aim to verify whether a suitable trade-off between accuracy and carbon footprint can be made, as justified in previous studies [31, 32]. By using different types of models with different sizes and complexity, a comprehensive analysis of accuracy and environmental sustainability for AI used in X-ray diagnosis can be conducted.

3.1 Application

For this paper, the low-code platform Mendix is used to test the use of LLMs and small models in the analysis of X-ray images. Mendix is a low-code platform that facilitates rapid application development and has seen success in developing medical applications that integrate machine learning through API capabilities [38]. Additionally, Mendix provides seamless integration of machine learning models by using the ONNX format. ONNX has been observed to allow the straightforward implementation of trained models [39]. Additionally, Mendix provides access to their own generative AI tools, GenAI, which uses Anthropic Claude-3.5 Sonnet hosted on AWS. Using this tool in conjunction with the OpenAI API within the Mendix application, multiple LLM models can be used for X-ray analysis. Figure 2 shows how the models will be implemented and interacted with in the Mendix application. The process of uploading and viewing the output of models is identical throughout all models. This simplifies the testing process and shows that Covid-19 detection can be performed identically regardless of the underlying model being used.

Mendix relies predominately on Java, which is not a favourable language for machine learning research. Using Java for AI algorithms has been found to have reduced CPU efficiency compared to Python [40]. However, by implementing the models in a different environment, the robustness of their accuracy will be tested outside of the research-focused environment of Python. This will help determine whether the models are suitable in scenarios where they will be deployed in applications for hypothetical real-word use. Therefore, the local models (Covid-Net, DenseNet, ResNet, and VGG) will be tested instead of using their stated accuracies from previous research to ensure that their performance is guaranteed in an unfavourable coding environment.

3.2 Comparison

LLM responses can provide descriptive analysis on X-rays. However, this does not guarantee improved interpretability. Hallucination or bias responses based on input prompts can cause uncertainty in the suggested approximation of signs of Covid-19 being present on X-rays. This is seen in other research, where LLMs have shown confidence in outputs that were later shown to be erroneous [19, 7]. Instead, it can be argued that the output from the small models is more interpretable, since they will consistently provide a probability on signs of Covid-19 infection. This shows smaller models having improved interpretability and transparency in narrow scopes [22]. So, while LLMs can provide detailed analysis, this risks causing confusion to end users if the details are false.

By using local discriminative models and generative LLM models, the default output of models can be difficult to compare. Although LLMs can provide details on X-ray images, medical expertise is required to verify the accuracy

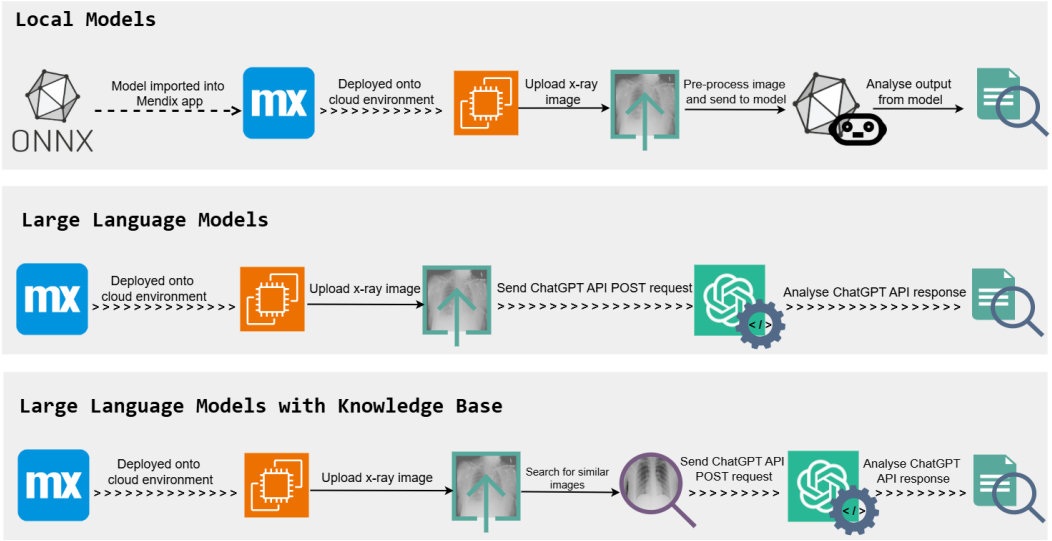


Figure 2: Deployment and use of models analysing X-ray images.

[10]. Furthermore, discriminative models only provide probabilities that Covid-19 symptoms are present. To allow for the straightforward comparison of LLMs and small models, output accuracy will be determined by the probability that an X-ray shows signs of symptoms associated with Covid-19. Therefore, LLMs will be provided with system instructions to limit outputs to Covid-19 probabilities to allow comparison with the outputs from local models.

To allow a straightforward comparison between models, the fine-tuning (where applicable) and testing models are performed with the Covidx CXR-3 dataset [28]. With labelled X-ray images in this open source dataset, the accuracy of models can be based on whether signs of Covid-19 are detected.

When analysing the time taken and carbon footprint of inferring models, control variables such as cloud hardware are important considerations. Comparing models on different servers brings uncertainty with different hardware being used [33]. By keeping the application and local models on the same server, more consistency can be guaranteed by ensuring the resources provided throughout testing models remains as consistent as possible.

With environmentally efficient AI-assisted applications being an important consideration in this paper, the carbon footprint for running the models is an important comparison. With the Mendix cloud using Kubernetes on AWS, Equation 1 can be used to estimate the carbon footprint of running a model on the deployed application. This equation is based on estimates of power consumption of EC2 instances on AWS [35]. To determine how long cloud resources are used, the time taken from uploading an X-ray to receiving probabilities is recorded. It is important to note that Kuberentes on AWS is priced every second for a minimum of 60 seconds, so the calculated carbon footprints are a hypothetical for the total runtime of the models, even if they run for less than 60 seconds.

Calculating the carbon footprint of LLMs hosted on servers is more challenging due to the variety of hardware used and unknown factors such as memory usage. However, information regarding the infrastructure specifications provided in existing research [37] can be used in Equation 1. This existing research has omitted manufacturing emissions to prevent a distortion in comparing models. Moreover, assuming the manufacturing emissions are shared across millions of daily inferences to the models over the lifespan of the hardware, the values should be negligible compared to the energy consumed through model inference. However, manufacturing emissions are not negligible for the hosting of the application, so these are still included in Table 1. So, while manufacturing emissions will be included in the calculation of the carbon footprint produced by the Mendix application and discriminative models, Table 1 shows that these emissions will be omitted when calculating the carbon footprint produced by the servers hosting the LLMs.

$$\frac{E}{1000} = \left(\frac{W \times P \times I}{1000} + M\right) \times \frac{t}{3600} \tag{1}$$

$$E = \left(\frac{W \times P \times I}{1000} + M\right) \times \frac{t}{3.6} \tag{2}$$

$$\begin{aligned}\text{where } E &= \text{Carbon footprint (mgCO}_2\text{eq)},\\ W &= \text{Instance power consumption (Watts)},\\ P &= \text{Power usage effectiveness},\\ I &= \text{Electricity carbon intensity (gCO}_2\text{eq/kWh)},\\ M &= \text{Manufacturing emissions (gCO}_2\text{eq)},\\ t &= \text{Time model is running (seconds)}\end{aligned}$$

It is important to note that the memory utilised for X-ray analysis is contingent upon the specific model used. With exact power consumption for specific memory used being unavailable, Equation 3 calculates the carbon footprint in the context of the amount of memory used in the cloud instance during X-ray analysis. When using LLMs, the size of the model (m) will be 0 as the model inference occurs through an API call rather than deploying the model into the application environment.

$$E_m = E \times \frac{a+m}{C} \tag{3}$$

$$\begin{aligned}\text{where } E_m &= \text{Carbon footprint per cloud memory used (mgCO}_2\text{eq/MB)},\\ E &= \text{Carbon footprint (mgCO}_2\text{eq)},\\ a &= \text{Application size (MB)},\\ m &= \text{Model size (MB)},\\ C &= \text{Cloud instance total memory (MB)}\end{aligned}$$

At the time of analysis, GPT-4.5-Preview was the most recent release provided by the OpenAI API. Other studies have used ChatGPT extensions for X-ray diagnosis, with the "X-Ray Interpreter" being a notable example [11]. However, there was no identifiable way to use these extensions through API calls to ChatGPT through the Mendix application. Moreover, initial tests using the extension showed no discrepancy in the diagnosis probabilities outputted compared to using GPT-4.5.

Table 1: Infrastructure specifications of instances.

Instance	Instance power consumption (W)	PUE	Carbon intensity (gCO_eq/kWh)	Manufacturing emissions (gCO_2eq)
App	5.3	1.2	228	1.2
GPT-4.5-Preview	1301	1.12	353	na
o4-Mini	991	1.12	353	na
GPT-4.1-Nano	377	1.12	353	na
GenAI	1301	1.14	385	na

To verify whether the size of the model affected the performance in X-ray analysis, smaller variations of GPT models were also used. Therefore, analysis of the test dataset was conducted using the OpenAI API to send requests to GPT-4.5-Preview, o4-Mini, and GPT-4.1-Nano. Alongside Mendix's GenAI being analysed, this allows the benchmark study to compare five LLMs and four discriminative models.

3.3 Refinement with knowledge bases

After initial analysis with the LLMs, providing additional information was necessary due to low accuracy. Knowledge bases have been shown to improve medical image analysis with LLMs [18]. However, exposing additional sensitive information to LLMs can threaten patient safety by increasing the risk of malpractice [5]. Additionally, providing additional images would increase the cost of using LLMs dramatically (providing three similar images alongside the original could hypothetically quadruple the input tokens used). Therefore, a knowledge base containing medical images was created, whereby the similarity of images to the uploaded X-ray is compared instead of providing additional images to the LLMs. A cosine similarity search is used to retrieve the three most similar image vectors from the knowledge base, with this technique successfully improving LLM medical analysis [17]. The cosine similarity score

and whether the image was a positive or negative case will be used to provide additional information to the LLM. This allows a knowledge base to be integrated into the medical application without exposing additional medical images to third-party LLMs.

To store images as vectors in the knowledge base, the images need to be embedded. With the local models used in this medical application, their infrastructure can be modified to use their classification layer as the output to be vectorised. This allows the X-ray images to be put into a tensor format which can be vectorised. With Covid-Net and DenseNet having high accuracies in detecting Covid-19 on the chest X-rays, these models were used as classifiers for two separate knowledge bases. These knowledge bases consist of classification vectors for the X-ray images contained in the training section of the Covidx CXR-3 dataset. Although VGG also demonstrated high accuracy, using its classification layer resulted in a model that is five times greater in size than the other local models (500+ MB). To prevent the application from running out of memory, VGG was not chosen to vectorise images.

During the process of analysing LLMs with knowledge bases, the GPT-4.5-Preview was removed from the OpenAI API. Although this restricts the analysis that can be done with this model, it is still included in the results of this research to provide insight into its accuracy and environmental impact.

4 Results and discussion

4.1 Performance

Table 2: Performance of models when tested on Covidx CXR-3 dataset.

	Accuracy (%)	Median time taken (ms)	Median carbon footprint ($mgCO_2eq$)	Total memory used (%)	Median carbon footprint per cloud memory used ($mgCO_2eq/MB$)
Local models					
Covid-Net	**95.5**	907	0.668	42.5	0.284
DenseNet	91.8	264	0.194	40.5	0.0788
ResNet	82.3	**174**	**0.141**	41.6	**0.0588**
VGG	93.8	221	0.174	39.7	0.0693
LLMs					
GPT-4.5-Preview	52.3	7270	974	**36.8**	359
o4-Mini	47.5	5280	543	**36.8**	200
GPT-4.1-Nano	52.8	1670	59.4	**36.8**	21.9
GenAI	48.5	1580	220	37.9	85.5
LLMs with DenseNet knowledge base					
o4-Mini	60.5	5440	548	41.2	226
GPT-4.1-Nano	79.8	1730	56.4	41.2	23.3
GenAI	73.5	1890	236	42.4	100
LLMs with Covid-Net knowledge base					
o4-Mini	51.8	6460	565	43.3	245
GPT-4.1-Nano	79.3	3070	73.8	43.3	32.0
GenAI	73.8	2700	249	44.5	111

Apart from Covid-Net, local models are more likely to diagnose X-rays as Covid-19 positive rather than negative. This is shown in Table 3, where the sensitivity of local models, except Covid-Net, is significantly higher than their specificity. Meanwhile, the LLMs mostly follow Covid-Net with higher specificity values, regardless of whether a knowledge base is used. These LLM results, notably GenAI, are similar to existing research which found that the Claude-3.5 Sonnet model marked a high number of positive cases as normal (false negative) [4]. Reducing false negatives is important in medical diagnosis as they can risk delayed or missed treatment [9]. Therefore, it can be argued that the bias towards positive diagnosis from most models is safer than allowing false negatives. However, this bias means that these models should be used as supplementary tools rather than automating diagnosis, especially LLMs and ResNet with their lower accuracies.

When considering both accuracy and carbon footprint, VGG appears to be an appropriate solution to detect Covid-19 on chest X-rays. Table 2 shows that although the accuracy of VGG is 1.7% lower than that of Covid-Net, it also has a 74% reduced carbon footprint. With this accuracy being higher than existing models [27, 29], it could be argued that this accuracy is sufficient with the reduction in carbon footprint. However, Table 3 shows that the positive predictive value (PPV) of VGG is 10% lower than that of Covid-Net. This highlights a significantly higher rate of false positives, which could waste time and resources due to incorrect diagnoses. So, while VGG significantly reduces the carbon footprint, the higher risk of incorrect diagnosis means that this may not be an appropriate trade-off.

Table 3: Accuracy, specificity, sensitivity, and positive predictive value (PPV) of models.

	Accuracy (%)	Specificity (%)	Sensitivity (%)	PPV (%)
Local models				
Covid-Net	**95.5**	**99.0**	92.0	**98.9**
DenseNet	91.8	84.5	99.0	86.5
ResNet	82.3	64.5	**100**	73.8
VGG	93.8	87.5	**100**	88.9
LLMs				
GPT-4.5-Preview	52.3	48.0	56.5	52.1
o4-Mini	47.5	60.5	34.5	46.6
GPT-4.1-Nano	52.8	70.5	35.0	54.3
GenAI	48.5	48.5	48.5	48.5
LLMs with DenseNet knowledge base				
o4-Mini	60.5	82.5	38.5	68.8
GPT-4.1-Nano	79.8	84.5	75.0	82.9
GenAI	73.5	96.0	51.0	92.7
LLMs with Covid-Net knowledge base				
o4-Mini	51.8	78.0	25.5	53.7
GPT-4.1-Nano	79.3	83.5	75.0	82.0
GenAI	73.8	88.0	59.5	83.2

Receiving a probabilistic output from LLMs proved to be an initial challenge. This could be censorship where the LLMs are trying to prevent the disclosure of sensitive information with the risk that malicious actors may misuse it [13]. Likewise, reluctance to output only probabilities could be because LLMS are better at providing descriptive analysis of X-rays rather than quantifiable probabilities of diseases. This is evident in Table 4, where reassurance in the input prompt that the results will be verified by a human radiologist is required for the model to return the predicted probabilities. As evident in Table 3, LLMs struggle to give accurate probabilities of Covid-19 being present without knowledge bases. This goes against existing work that notes ChatGPT provides accurate X-ray descriptions [10, 11]. With this paper restricting the output scope to not include these descriptions, ChatGPT struggled to give definitive diagnoses to X-rays.

Analysis of X-rays using LLMs appears to be biased towards the instructions initially given to the model. As observed in Figure 3, the estimated probability changes with an alteration in the prompt. This bias was further observed when the API request was improperly structured, as noted in A, whereby when an image was not provided within the prompt, the response would skew to whether the prompt asks for a positive or negative diagnosis. Although the LLM is tasked with analysing signs of Covid-19 in all three scenarios, the output is skewed to the underlying tones in the prompt; asking for probability of no Covid-19 symptoms pushes a bias that the result is expected to be negative. These results are in line with the suggestions of existing works to use LLM results as a supplementary tool rather than a replacement for patient communication [3]. So, while LLMs can provide greater detail than smaller models, the significant risk of bias results means the model should only be used as a supplementary tool, similar to that of the smaller models.

When knowledge bases are introduced, a significant increase to the accuracy is provided. The GPT-4.1-Nano model had a 27% increase in accuracy when a knowledge base with DenseNet classification is used. GenAI saw a 47.5% increase in specifity, making it only second to Covid-Net in identifying Covid-19 negative patients. The o4-Mini model was influenced the least by the knowledge bases, with the Covid-Net classification knowledge base producing an accuracy still lower than GPT-4.5-Preview had without a knowledge base. This highlights how some models benefit more from knowledge bases [17]. While introducing knowledge bases was found to increase LLM accuracy universally, both the type of LLM and classification model used in the knowledge base influence how impactful the additional information is when analysing X-rays.

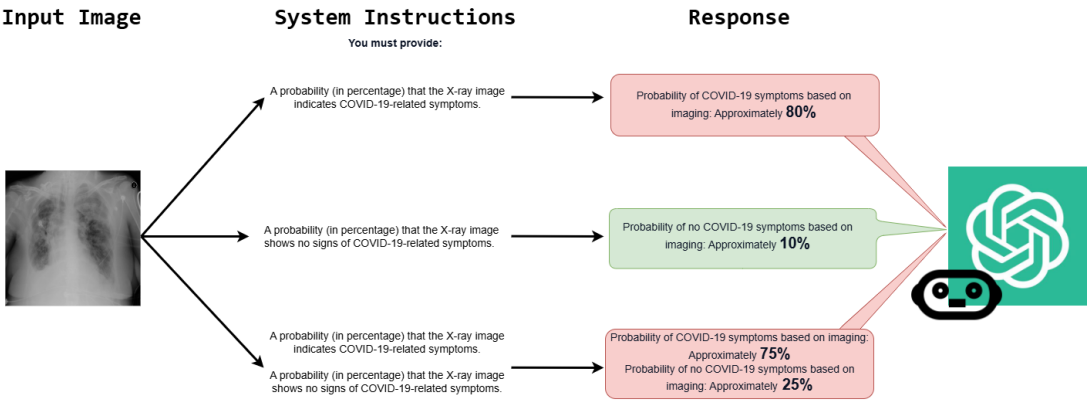


Figure 3: ChatGPT response alteration depending on system instructions.

Using LLMs without knowledge bases for Covid-19 detection resulted in poor accuracy. As shown in Figure 4, these models showed high confidence in their output, even when it was a false positive/negative. Meanwhile, they rarely outputted probabilities that suggested uncertainty in the provided X-ray (e.g., approximating an only 51% probability in an X-ray displaying signs of Covid-19). This highlights the hallucination risks of using ChatGPT to provide probabilistic outputs for Covid-19, whereby admission of uncertainty is rare.

Knowledge bases resulted in LLMs having more confidence in their analysis. With GenAI, correct analysis with high confidence of 90-100% increased from 0.01 to 0.26 when a knowledge base with Covid-Net embedding is used. Generally, Covid-Net embedding provided more confidence than the DenseNet embedding knowledge base. This can be seen with GPT-4.1-Nano, where the confidence range of 80-100% increases from 0.328 to 0.49 when using Covid-Net embedding over the DenseNet embedding knowledge base. Although GPT-4.1-Nano and GenAI have similar accuracies regardless of the knowledge base used, more confident analysis appears when using Covid-Net embedding.

The local models outperformed the accuracy of LLMs, even when knowledge bases were used. Notably, Covid-Net demonstrated high confidence in outputting accurate results; over 90% of its output correctly identified X-rays as positive or negative with a high confidence of 90-100%. Meanwhile, the other discriminative models only reach this accuracy when increasing the confidence range to 70-100%. Therefore, while models such as DenseNet and VGG have an accuracy of over 90%, the confidence in their outputs is less than that of Covid-Net, suggesting a higher level of doubt in the analysis when using these models. This highlights the superiority of Covid-Net in confidently identifying Covid-19 symptoms compared to existing model architecture, which was fine-tuned to work in the scope outlined in this paper.

4.2 Environmental impact

When comparing the locally deployed models, Covid-Net shows a slower response time and consequently a greater carbon footprint. As shown in Figures 5a and 6, the efficiency of Covid-Net is around 3.5 times less than that of the other local models. While this makes Covid-Net less environmentally efficient in the context of the other local models, its impact is considerably less than the impact from GPT-4.5-Preview. Figures 5b and 6 show that GPT-4.5-Preview has an exponentially greater environmental impact. The average time taken for GPT-4.5-Preview is 8 times greater than using Covid-Net. Even when taking into consideration the reduced application size of using API calls to LLMs, Figure 7 shows GPT-4.5-Preview still results in a greater footprint a thousand times greater than Covid-Net when analysing the X-ray images. So, while Covid-Net is environmentally inefficient compared to other local models, it still produces a smaller carbon footprint compared to LLMs.

GenAI demonstrated the fastest and most consistent speed across the LLMs. When no knowledge base is used, its mean of 1730ms and interquartile range (IQR) of 253ms is not far off the mean and IQR of Covid-Net (when compared to the other LLMs) which were 929ms and 117ms respectively. Due to this faster response time, GenAI achieves a median footprint 80% lower than GPT-4.5-Preview despite having similar server hardware and GPU utilisation. Although the speed of GenAI is close to local models, Figure 6 shows a significant disparity in their carbon footprint produced. The difference in carbon footprint is further explored in B, highlighting that the smaller local models have a 99% reduced carbon footprint compared to GPT-4.5-Preview and GenAI if these models are continuously run for a longer period of time. So, even when LLM response times are fast, the interaction with a server hosting the LLM results in a significant

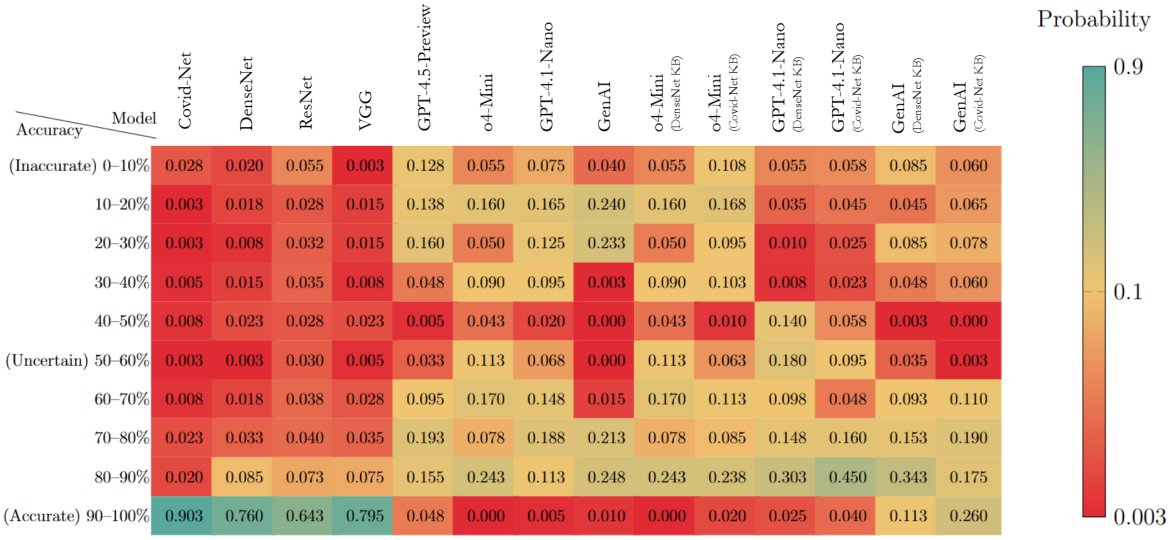


Figure 4: Distribution of model output accuracy, where 100% shows high accuracy and 0% shows high inaccuracy (false positive/negative).

increase in the carbon footprint, highlighting the adverse effect on energy consumption when using large models for the classification task.

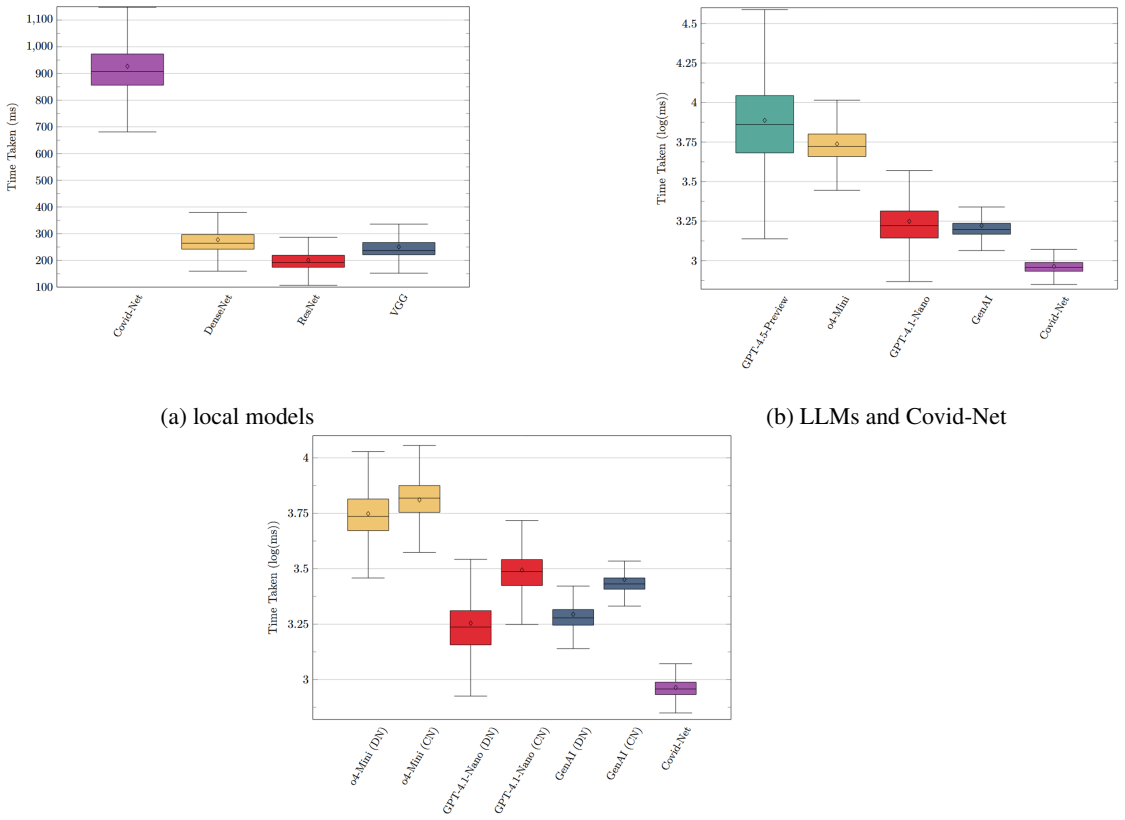
While Covid-Net demonstrates superior accuracy, it could be argued that the VGG model provides a more competitive solution because of its reduced carbon footprint. This would support existing research, where a slight reduction in accuracy is accepted for reduced carbon emissions [31, 32]. However, Figure 4 highlights VGG has less confidence in its outputs, which could cause doubt in diagnosis. So, while models smaller than Covid-Net have a reduced carbon footprint, the reduced confidence in output makes these models less confident due to risking doubt in Covid-19 diagnosis.

The carbon footprint from using LLMs highlights the more energy-intensive process of LLMs compared to smaller neural networks. A reason for this energy consumption is that models that generate new text are more carbon intensive [36]. Although the response from LLMs has been narrowed to only return probabilities in the output, the text it returns is still newly generated content. Meanwhile, the small neural networks are constrained to binary classification, outputting two decimal values instead of generating new text. This shows that even if models are given the same narrow scope to complete a task, the type of model and the way it formats tasks have a significant role in the carbon footprint produced for an application utilising that model.

Introducing knowledge bases saw a mixed impact in carbon footprint. GPT-4.1-Nano saw a 24.2% increase when using the Covid-Net embedded knowledge base. Meanwhile, o4-Mini and GenAI only saw a 4.0% 13.1% increase in carbon footprints respectively with this knowledge base. This could be because GPT-4.1-Nano has a lower carbon footprint than the other LLMs, meaning a change in its carbon footprint is more noticeable. Meanwhile, using the DenseNet embedded knowledge base saw a decrease of 5.0% in carbon footprint for GPT-4.1-Nano; while the process of uploading an image and waiting for an output was longer, the time for the LLM running was shorter, which is where the majority of the carbon footprint is being produced. The o4-Mini and GenAI models saw increases of 1% and 7.2% respectively with the DenseNet embedded knowledge base. This shows that the environmental impact of using knowledge bases with LLMs varies across the different models and how information is embedded.

5 Future work and limitation

With uncertainties surrounding the energy consumption when training models, the calculated carbon footprint produced by the application was constrained to inference with the models. This is a considerable limitation of this paper, whereby the environmental sustainability of training LLMs and local models is not tested. It is recommended that future work investigate the training stage of the models employed in X-ray analysis and whether the collaborative nature of third-party LLMs could demonstrate greater environmental sustainability than that implied through the model inference of the medical application in this paper.



(a) local models

(b) LLMs and Covid-Net

(c) LLMs with DenseNet (DN) and Covid-Net (CN) knowledge bases

Figure 5: Comparing (a) time taken (b), (c) log time taken for models to return output from an X-ray image input.

At the time of analysis of the models, there was no suitable way of fine-tuning ChatGPT models with images, nor a way to use ChatGPT extensions within Mendix. This gave ChatGPT a disadvantage in analysis, whereby it may have never been trained on the training dataset the local models were exposed to. Future work should explore the capabilities of fine-tuning LLMs like ChatGPT and analysing their accuracy to local models.

The scope of this paper focused on classifying chest X-rays as either Covid-19 positive or negative. This binary classification allowed for straightforward comparison between models, which would typically be a challenge due to generative and discriminative models producing different outputs. Consequently, this research does not provide analysis on wider, multi-class scenarios such as identifying different types of diseases from chest X-rays. Future work should determine whether the superior accuracy and environmental efficiency of local discriminative models in binary classification can also outperform generative models in multi-classification scenarios.

6 Conclusion

Medical applications have benefitted from integration with AI tools. However, greater scrutiny is being brought to their accuracy and environmental impact. With the versatility of LLMs, generative models are sometimes used in classification tasks despite showing a reduced comprehension in these scenarios [21]. In contrast, small discriminative models can outperform the accuracy of LLMs when used in case-specific applications [23]. However, reducing model size to reduce environmental impact can adversely affect accuracy. This paper investigates the performance of third-party LLMs and small local models in detecting Covid-19 in chest X-rays. ResNet and VGG resulted in the application having a lower carbon footprint, but their lower PPV values highlighted a strong bias to false positives. Although Covid-Net results in a carbon footprint 3.5 times greater than these models, its footprint is still a 99.93% reduction than that produced when using GPT-4.5-Preview. Even more environmentally friendly models such as GPT-4.1-Nano, which has a 94.2% lower carbon footprint than GPT-4.5-Preview, still produces a footprint 8350% greater than Covid-Net. Moreover, Covid-Net had an accuracy 15.7% higher than that of the best performing LLM (GPT-4.1-Nano with DenseNet embedded knowledge base). This highlights the risk of hallucinations when using LLMs for

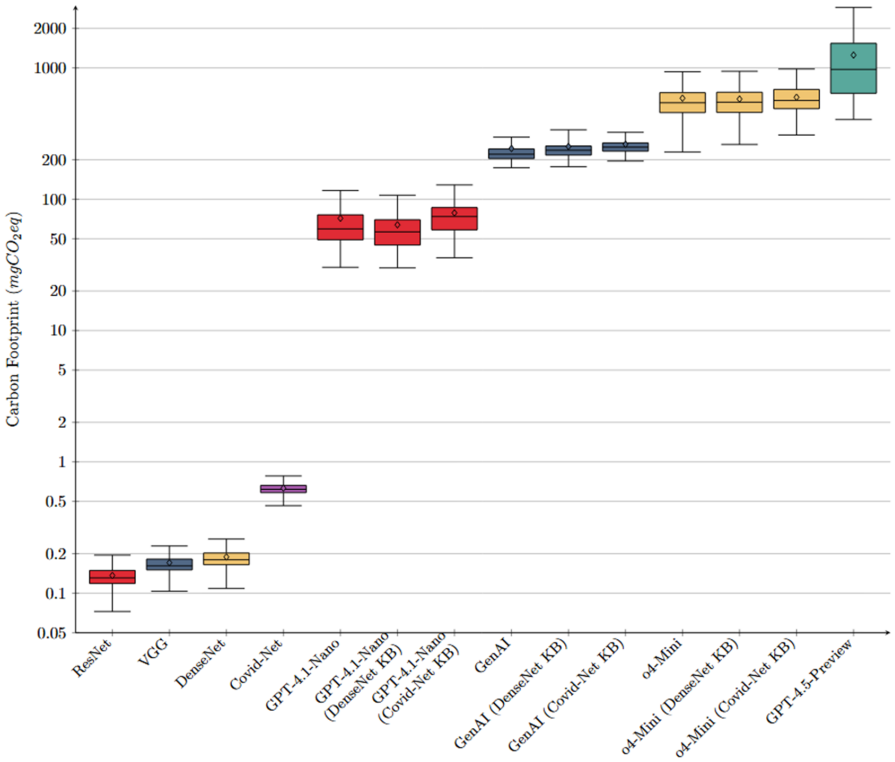


Figure 6: Comparing carbon footprint produced by models to return output from an X-ray image input.

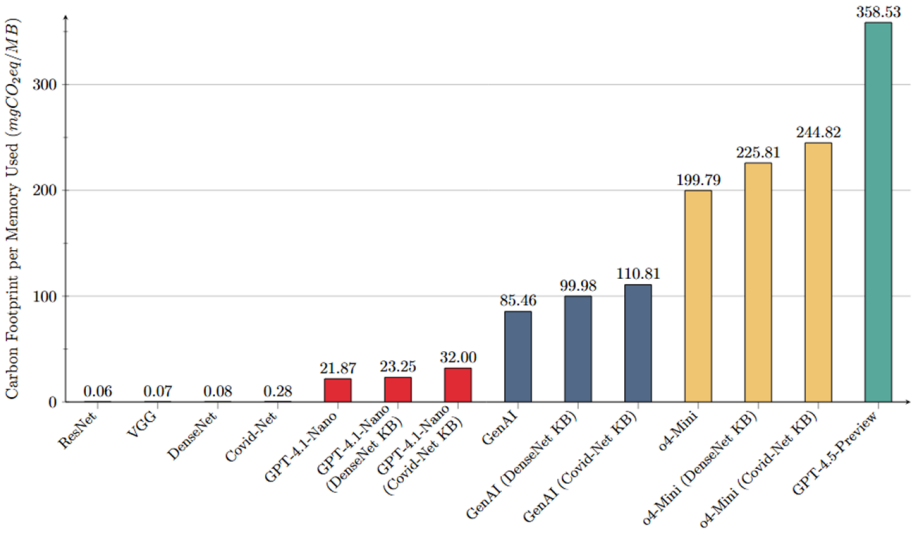


Figure 7: Comparing carbon footprint per memory used produced by models to return output from an X-ray image input.

medical diagnosis [19]. These results show that larger models used in a classification task for Covid-19 detection have a disproportionate carbon footprint and sub-optimal accuracy when compared to small discriminative model. Additionally, discriminative model sizes should not be reduced to an absolute minimum in an attempt to sacrifice accuracy for the lowest possible carbon footprint, especially since their increased carbon footprint is negligible when compared to the power consumption of LLMs. Future work should investigate the carbon footprint associated with training these models used in medical applications and whether discriminative models can continue outperforming generative models in multi-class scenarios.

References

[1] M. U. Hadi, Q. Al-Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects," 5 2023.

[2] V. Sorin, E. Klang, M. Sklair-Levy, I. Cohen, D. B. Zippel, N. B. Lahat, E. Konen, and Y. Barash, "Large language model (chatgpt) as a support tool for breast tumor board," *npj Breast Cancer*, vol. 9, 2023.

[3] P. Keshavarz, S. Bagherieh, S. A. Nabipoorashrafi, H. Chalian, A. A. Rahsepar, G. H. J. Kim, C. Hassani, S. S. Raman, and A. Bedayat, "Chatgpt in radiology: A systematic review of performance, pitfalls, and future perspectives," *Diagnostic and Interventional Imaging*, vol. 105, pp. 251–265, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211568424001050

[4] C. Nguyen, D. Carrion, and M. K. Badawy, "Comparative performance of claude and gpt models in basic radiological imaging tasks," *medRxiv*, 2024. [Online]. Available: https://www.medrxiv.org/content/early/2024/11/18/2024.11.16.24317414.1

[5] D. O. Shumway and H. J. Hartman, "Medical malpractice liability in large language model artificial intelligence: Legal review and policy recommendations," 2024.

[6] S. Chen, X. Li, M. Zhang, E. H. Jiang, Q. Zeng, and C.-H. Yu, "Cares: Comprehensive evaluation of safety and adversarial robustness in medical llms," 2025. [Online]. Available: https://arxiv.org/abs/2505.11413

[7] M. Nezhurina, L. Cipolina-Kun, M. Cherti, and J. Jitsev, "Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models," 2025. [Online]. Available: https://arxiv.org/abs/2406.02061

[8] F. X. Doo, J. Vosshenrich, T. S. Cook, L. Moy, E. P. Almeida, S. A. Woolen, J. W. Gichoya, T. Heye, and K. Hanneman, "Environmental sustainability and ai in radiology: A double-edged sword," 2024.

[9] T. Noguchi, K. Yamashita, S. Matsuura, R. Kamei, J. Maehara, K. Furuya, S. Harada, S. Adachi, and Y. Okada, "Analysis of "visible in retrospect" to monitor false-negative findings in radiological reports," *Japanese Journal of Radiology*, vol. 41, 2023.

[10] O. Thawakar, A. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. Khan, J. Laaksonen, and F. S. Khan, "Xraygpt: Chest radiographs summarization using medical vision-language models," 2025. [Online]. Available: https://arxiv.org/abs/2306.07971

[11] A. M. Ostrovsky, "Evaluating a large language model's accuracy in chest x-ray interpretation for acute thoracic conditions," *The American Journal of Emergency Medicine*, vol. 93, pp. 99–102, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0735675725002256

[12] M. Ahmed and J. Knockel, "The impact of online censorship on llms," *Free and Open Communications on the Internet*, 2024.

[13] D. Glukhov, I. Shumailov, Y. Gal, N. Papernot, and V. Papyan, "Llm censorship: A machine learning challenge or a computer security problem?" 2023. [Online]. Available: https://arxiv.org/abs/2307.10719

[14] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu, "Prompt injection attack against llm-integrated applications," 2024. [Online]. Available: https://arxiv.org/abs/2306.05499

[15] O. Aydin and E. Karaarslan, "Openai chatgpt interprets radiological images: Gpt-4 as a medical doctor for a fast check-up," *arXiv preprint arXiv:2501.06269*, 2025.

[16] M. Ranjit, G. Ganapathy, R. Manuel, and T. Ganu, "Retrieval augmented chest x-ray report generation using openai gpt models," in *Proceedings of Machine Learning Research*, vol. 219, 2023.

[17] S. T. Arasteh, M. Lotfinia, K. Bressem, R. Siepmann, L. Adams, D. Ferber, C. Kuhl, J. N. Kather, S. Nebelung, and D. Truhn, "Radiorag: Online retrieval–augmented generation for radiology question answering," *Radiology: Artificial Intelligence*, vol. 7, 7 2025. [Online]. Available: http://dx.doi.org/10.1148/ryai.240476

[18] Z. Zhao, S. Wang, J. Gu, Y. Zhu, L. Mei, Z. Zhuang, Z. Cui, Q. Wang, and D. Shen, "Chatcad+: Toward a universal and reliable interactive cad using llms," *IEEE Transactions on Medical Imaging*, vol. 43, pp. 3755–3766, 2024.

[19] K. Bera, A. Gupta, S. Jiang, S. Berlin, N. Faraji, C. Tippareddy, I. Chiong, R. Jones, O. Nemer, A. Nayate, S. H. Tirumani, and N. Ramaiya, "Assessing performance of multimodal chatgpt-4 on an image based radiology board-style examination: An exploratory study," *medRxiv*, 2024. [Online]. Available: https://www.medrxiv.org/content/early/2024/01/13/2024.01.12.24301222

[20] A. Heiman, X. Zhang, E. Chen, S. E. Kim, and P. Rajpurkar, "Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models," 2024. [Online]. Available: https://arxiv.org/abs/2411.18672

[21] H. Xu, R. Lou, J. Du, V. Mahzoon, E. Talebianaraki, Z. Zhou, E. Garrison, S. Vucetic, and W. Yin, "Llms' classification performance is overclaimed," 2024. [Online]. Available: https://arxiv.org/abs/2406.16203

[22] L. Chen and G. Varoquaux, "What is the role of small models in the llm era: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2409.06857

[23] M. Hassid, T. Remez, J. Gehring, R. Schwartz, and Y. Adi, "The larger the better? improved llm code-generation via budget reallocation," 2024. [Online]. Available: https://arxiv.org/abs/2404.00725

[24] S. Sharma and K. Guleria, "A deep learning based model for the detection of pneumonia from chest x-ray images using vgg-16 and neural networks," in *Procedia Computer Science*, vol. 218, 2022.

[25] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," 2021.

[26] S. Bharati, P. Podder, M. R. H. Mondal, and V. B. Prasath, "Co-resnet: Optimized resnet model for covid-19 diagnosis from x-ray images," *International Journal of Hybrid Intelligent Systems*, vol. 17, 2021.

[27] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, p. 19549, 11 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-76550-z

[28] M. Pavlova, T. Tuinstra, H. Aboutalebi, A. Zhao, H. Gunraj, and A. Wong, "Covidx cxr-3: A large-scale, open-source benchmark dataset of chest x-ray images for computer-aided covid-19 diagnostics," 2022. [Online]. Available: https://arxiv.org/abs/2206.03671

[29] S. Albahli, N. Ayub, and M. Shiraz, "Coronavirus disease (covid-19) detection using x-ray images and enhanced densenet," *Applied Soft Computing*, vol. 110, 2021.

[30] L. Lannelongue, J. Grealey, and M. Inouye, "Green algorithms: Quantifying the carbon footprint of computation," *Advanced Science*, vol. 8, 2021.

[31] J. Jung, "Carbon intensity-aware adaptive inference of dnns," 2024. [Online]. Available: https://arxiv.org/abs/2403.15824

[32] S. Narimani, S. R. Hoff, K. D. Kurz, K.-I. Gjesdal, J. Geisler, and E. Grøvik, "Comparative analysis of deep learning architectures for breast region segmentation with a novel breast boundary proposal," *Scientific Reports*, vol. 15, p. 8806, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-92863-3

[33] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of bloom, a 176b parameter language model," *Journal of Machine Learning Research*, vol. 24, pp. 1–15, 2023. [Online]. Available: http://jmlr.org/papers/v24/23-0069.html

[34] E. Barbierato and A. Gatti, "Toward green ai: A methodological survey of the scientific literature," *IEEE Access*, vol. 12, 2024.

[35] B. Davy, "Building an aws ec2 carbon emissions dataset," 9 2021. [Online]. Available: https://medium.com/teads-engineering/building-an-aws-ec2-carbon-emissions-dataset-3f0fd76c98ac

[36] S. Luccioni, Y. Jernite, and E. Strubell, "Power hungry processing: Watts driving the cost of ai deployment?" in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2024, pp. 85–99. [Online]. Available: https://doi.org/10.1145/3630106.3658542

[37] N. Jegham, M. Abdelatti, L. Elmoubarki, and A. Hendawi, "How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference," 2025. [Online]. Available: https://arxiv.org/abs/2505.09598

[38] A. Agarwal and A. Das, "An integrated breast cancer predictor leveraging mendix and machine learning," in *2025 6th International Conference on Recent Advances in Information Technology (RAIT)*, 2025, pp. 1–6.

[39] A. Shridhar, P. Tomson, and M. Innes, "Interoperating deep learning models with onnx.jl," *JuliaCon Proceedings*, vol. 1, 2020.

[40] M. M. A. Acevedo, "Java v.s. python in ai," Ph.D. dissertation, UPC, Facultat d'Informàtica de Barcelona, Departament d'Arquitectura de Computadors, 10 2021. [Online]. Available: http://hdl.handle.net/2117/361296

[41] N. Express, "Our commitment to the environment," 2025. [Online]. Available: https://www.nationalexpress.com/en/environment

A LLM hallucination

During initial development of the medical application, the OpenAI API request was formatted incorrectly. Consequently, no image was sent with the request when asking for analysis on the X-ray. Although no image was sent, the GPT model still returned a probability. With probabilities being returned by the LLMs, no issues were initially identified. When investigating the tokens used for the API requests, it was apparent that the images were not being sent. As shown in Table 4, the number of prompt tokens when an image is not sent is significantly smaller than when an image is sent. Moreover, the probability given when no image is supplied is drastically different to that when an image is provided. This highlights the risks of implementing ChatGPT into applications, whereby it is not initially evident that all essential information is not being supplied to the LLM.

Table 4: Prompt information for requesting analysis from GPT-4.5-Preview of an X-ray image.

Input prompt	Response	Prompt tokens	Image sent
You are an assistant radiologist, ready to dissect medical cases and give medical insights. You must give a probability that the provided x-ray shows Covid-19 symptoms. Assume that any recommendations you give will then be verified by a human radiologist.	There is approximately an **85–90%** probability that this X-ray is indicative of COVID-19 pneumonia.	66	No
You are an assistant radiologist, ready to dissect medical cases and give medical insights. You must give a probability that the provided x-ray shows Covid-19 symptoms. You must also give a probability that the provided x-ray shows someone with no Covid-19 symptoms. Assume that any recommendations you give will then be verified by a human radiologist. Output in the format: Probability of Covid-19 symptoms based on imaging: Approximately __ Probability of no Covid-19 symptoms based on imaging: Approximately __	Probability of Covid-19 symptoms based on imaging: Approximately **15%** Probability of no Covid-19 symptoms based on imaging: Approximately **85%**	366	Yes
You are an assistant radiologist trained in interpreting chest X-ray images for signs of COVID-19. Your task is to analyze the provided chest X-ray and estimate the likelihood that it shows signs consistent with COVID-19 infection. You must provide: A probability (in percentage) that the X-ray image indicates COVID-19-related symptoms (e.g., ground-glass opacities, bilateral infiltrates, or consolidation). A probability (in percentage) that the X-ray image shows no signs of COVID-19-related symptoms. Assume your analysis will be reviewed by a qualified human radiologist before informing any clinical decisions. Base your assessment solely on the imaging evidence. Output format: Probability of COVID-19 symptoms based on imaging: Approximately __% Probability of no COVID-19 symptoms based on imaging: Approximately __%	Probability of COVID-19 symptoms based on imaging: Approximately **10%** Probability of no COVID-19 symptoms based on imaging: Approximately **90%**	429	Yes

B Carbon footprint comparison

To provide context to the carbon footprint produced by the models, comparison with the carbon footprint produced by using an electric fan, electric heater, and a coach journey are made. The calculation of the carbon footprint from coach travel is based on the details provided by National Express - their Levante III coach produces 21.7 grams of greenhouse gas emission per kilometre travelled at 100% passenger capacity [41]. Using Equation 2, an estimation of the carbon footprint produced by the models over the travel time of the coach (assumed to be 3 hours) can be made.

Figure 8 demonstrates the vast differences in carbon footprint produced by the models used to analyse x-ray images. While using the small local models produces a carbon footprint 88% less than an electric fan, GPT-4.1-Nano produces the same carbon footprint as almost 6 electric fans. Using GenAI to analyse x-ray images over the same time period of a coach journey from Birmingham to London produces a carbon footprint 42% the size of the carbon footprint by that journey. Using the small models produces a carbon footprint 0.2% the size of the coach journey.

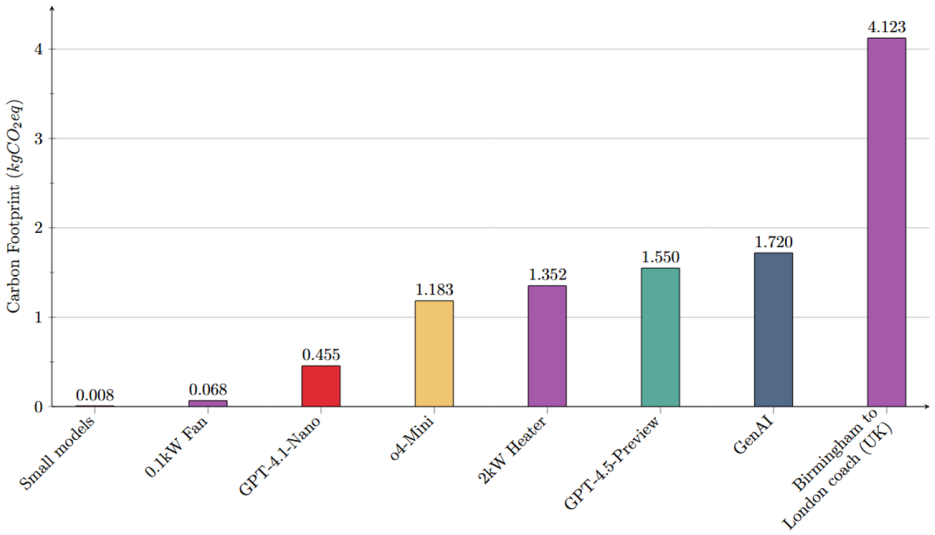


Figure 8: Carbon footprint produced over 3 hours by models, appliances, and coach travel.

Although GenAI has the highest carbon footprint out of the LLMs in this paper, it is important to note that it also had the quickest response time out of the LLMs. So, while GenAI has a carbon footprint 11% greater than GPT-4.5-Preview over a 3 hour period, it responded 4.6 times quicker when analysing X-ray images.