Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

Optimal Large Language Model Characteristics to Balance Accuracy and Energy Use for Sustainable Medical Applications

*Florence X. Doo, MD, MA** • *Dharmam Savani, BE** • *Adway Kanhere, MSE* • *Ruth C. Carlos, MD, MS* • *Anupam Joshi, PhD* • *Paul H. Yi, MD* • *Vishwa S. Parekh, PhD*

From the University of Maryland Medical Intelligent Imaging (UM2ii) Center, Department of Radiology and Nuclear Medicine, University of Maryland School of Medicine, 22 S Greene St, Baltimore, MD 21201 (F.X.D., D.S., A.K., P.H.Y., V.S.P.); Department of Radiology, University of Michigan, Ann Arbor, Mich (R.C.C.); and Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, Md (A.J.). Received February 1, 2024; revision requested February 23; revision received June 17; accepted June 18. **Address correspondence to** V.S.P. (email: *vparekh@som.umaryland.edu*).

* F.X.D. and D.S. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

Radiology 2024; 312(2):e240320 • https://doi.org/10.1148/radiol.240320 • Content codes: CH AI

Background: Large language models (LLMs) for medical applications use unknown amounts of energy, which contribute to the overall carbon footprint of the health care system.

Purpose: To investigate the tradeoffs between accuracy and energy use when using different LLM types and sizes for medical applications.

Materials and Methods: This retrospective study evaluated five different billion (B)–parameter sizes of two open-source LLMs (Meta's Llama 2, a general-purpose model, and LMSYS Org's Vicuna 1.5, a specialized fine-tuned model) using chest radiograph reports from the National Library of Medicine's Indiana University Chest X-ray Collection. Reports with missing demographic information and missing or blank files were excluded. Models were run on local compute clusters with visual computing graphic processing units. A single-task prompt explained clinical terminology and instructed each model to confirm the presence or absence of each of the 13 CheXpert disease labels. Energy use (in kilowatt-hours) was measured using an open-source tool. Accuracy was assessed with 13 CheXpert reference standard labels for diagnostic findings on chest radiographs, where overall accuracy was the mean of individual accuracies of all 13 labels. Efficiency ratios (accuracy per kilowatt-hour) were calculated for each model type and size.

Results: A total of 3665 chest radiograph reports were evaluated. The Vicuna 1.5 7B and 13B models had higher efficiency ratios (737.28 and 331.40, respectively) and higher overall labeling accuracy (93.83% [3438.69 of 3665 reports] and 93.65% [3432.38 of 3665 reports], respectively) than that of the Llama 2 models (7B: efficiency ratio of 13.39, accuracy of 7.91% [289.76 of 3665 reports]; 13B: efficiency ratio of 40.90, accuracy of 74.08% [2715.15 of 3665 reports]; 70B: efficiency ratio of 22.30, accuracy of 92.70% [3397.38 of 3665 reports]). Vicuna 1.5 7B had the highest efficiency ratio (737.28 vs 13.39 for Llama 2 7B). The larger Llama 2 70B model used more than seven times the energy of its 7B counterpart (4.16 kWh vs 0.59 kWh) with low overall accuracy, resulting in an efficiency ratio of only 22.30.

Conclusion: Smaller fine-tuned LLMs were more sustainable than larger general-purpose LLMs, using less energy without compromising accuracy, highlighting the importance of LLM selection for medical applications.

Supplemental material is available for this article.

The increasing adoption of energy-intensive artificial intelligence (AI), including large language models (LLMs), contributes to carbon emissions and exacerbates the climate crisis, with downstream health impacts (1–3). Early studies show promise in the ability of LLMs to accurately interpret medical language in clinical records such as radiology reports, which can help with patient summarization or image labeling tasks (4–6). However, the rising interest in medical applications of LLMs also raises concerns about the potential high energy use of LLMs.

Two fundamental LLM characteristics, model design and inherent size, may influence the balance of performance accuracy and energy use during medical applications. A general-purpose model is designed to handle a wide array of tasks without specific optimizations for any single domain, making it versatile across diverse applications, including radiology (7,8). In contrast, a specialized fine-tuned model may have specific enhanced capabilities to follow instructions, theoretically more suitable for clinical language instruction–based tasks (9,10).

LLM size is typically defined by the number of "parameters." The parameters in an LLM are akin to the weighted neurons in the human brain, where each one contributes to the model's overall knowledge and decision-making process. Therefore, the size of an LLM refers to its complexity and learning capacity such that more parameters mean the model can potentially recognize more nuanced patterns in the data, which could translate into higher accuracy for tasks such as diagnosing diseases from radiographs. Current models range in millions to trillions of parameters

Abbreviations

AI = artificial intelligence, GPU = graphic processing unit, LLM = large language model

Summary

This study explores the balance between accuracy and energy use in different-sized large language models for a medical application, highlighting the importance of model selection to balance performance with resource usage.

Key Results

- Using open-access data from 3665 chest radiograph reports, this study establishes a practical efficiency ratio (accuracy per kilowatt-hour) to evaluate the energy consumption of various large language model types and sizes.
- The 7-billion (B)–parameter Vicuna 1.5 (LMSYS Org) model had the highest efficiency ratio (737.28 vs 13.39 for the Llama 2 7B [Meta] model), while the Vicuna 7B and 13B models had the highest overall report-labeling accuracy at 93.83% and 93.65%, respectively.
- The larger Llama 2 70B model used more than seven times the energy of its 7B counterpart (4.16 kWh vs 0.59 kWh) with low overall accuracy, resulting in an efficiency ratio of only 22.30.

(11). However, this increased ability comes with a cost in terms of the energy required to process and analyze input data. The energy use related to the initial training and development of the models themselves can vary, ranging in megawatt-hours from 85.7 MWh for smaller LLMs, such as Google's T5 (11 billion parameters), to 1287 MWh for larger models, such as OpenAI's GPT-3 (175 billion parameters) (12,13). However, for clinical research and radiologic applications, most users will not train a new LLM from scratch but will leverage pretrained LLMs for inference tasks, which currently lack data on energy use.

Therefore, to enable sustainable AI choices, LLM clinical end users will need to understand the characteristics of LLMs that determine the judicious balance between accuracy and energy use, which currently remains a knowledge gap (14,15).

It is important to consider both potential model type and model size thresholds, beyond which further accuracy gains do not justify the associated energy costs. Thus, the aim of this study was to investigate the tradeoffs between accuracy and energy use when using a variety of LLM types and sizes for a medical application. In particular, five different sizes of two open-source LLMs (a general-purpose model and specialized fine-tuned model) were examined to identify 13 common disease labels within a publicly available chest radiographic data set.

Materials and Methods

This retrospective study used de-identified data and publicly available AI models and, thus, was exempt from institutional review board review. An overview of the study is shown in Figure 1, and the code is available on GitHub *(https://github.com/UM2ii/MedCrunchR)*.

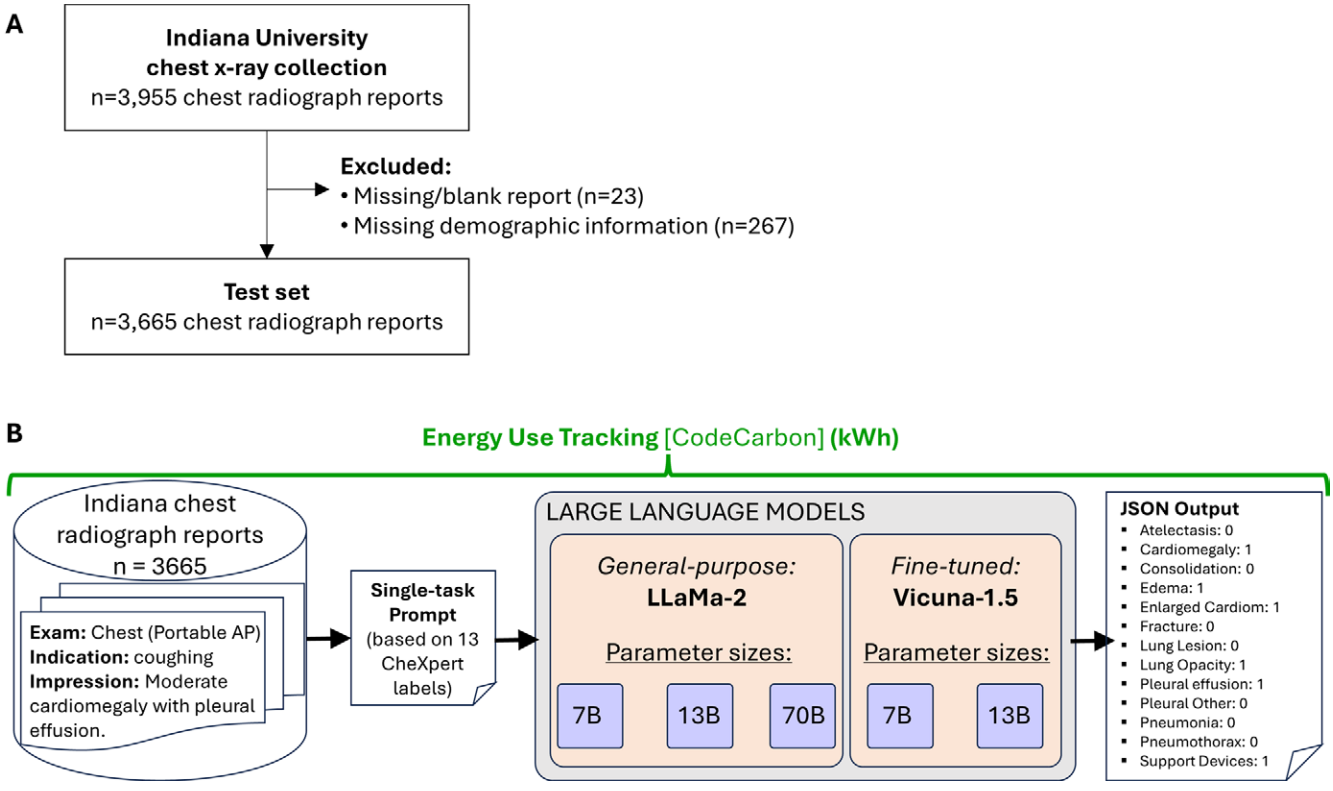

Figure 1: **(A)** Flowchart shows selection of the study data, which were derived from the Indiana University Chest X-ray Collection in the National Library of Medicine's Open-i and accessed on December 11, 2023. **(B)** Diagram shows an overview of the study. The publicly available open-source large language models, Meta's Llama 2 and LMSYS Org's Vicuna 1.5, were run locally using the chest radiographic data. CodeCarbon 2.3.1 is an open-source software tool designed to track energy use associated with computing tasks. Appendix S1 provides details on the prompt used. An output file was generated to report the results of 13 diagnostic findings according to CheXpert (Stanford ML Group) radiologist-labeled reference standard disease labels. AP = anteroposterior, B = billion, Cardiom = cardiomediastinum, JSON = JavaScript Object Notation.

LLMs and Parameter Sizes

To evaluate LLM efficiency (accuracy and energy use) for a prespecified medical labeling task, two LLMs were selected (Meta's Llama 2 [16] and LMSYS Org's Vicuna 1.5 [17]) based on three factors. First, both are open-source and publicly available models, and thus downloadable for local energy measurement and transparency. Second, both are established LLMs in their categories; Llama 2 is a popular general-purpose LLM (18) and Vicuna 1.5 is a model fine-tuned on Llama 2 that has enhanced capability to better interpret instructions (17,19). Lastly, each model has various default billion (B)–parameter sizes available (7B, 13B, and 70B for Llama 2; 7B and 13B for Vicuna 1.5). This range in model sizes and complexity (number of parameters) enabled a comparison of how energy use and performance may correlate with an increasing number of parameters. All of these base models were used with no additional adjustments or fine-tuning to ensure reproducibility and generalizability of study findings. Finally, the temperature hyperparameter was set to 0 for the experiments to reduce hallucinations and enhance consistency of the outputs for reproducibility.

Data Set Selection

This study used downloadable chest radiograph reports from the publicly available Indiana University Chest X-ray Collection (Indiana Network for Patient Care) available from the National Institutes of Health–National Library of Medicine's open-source biomedical image search engine, Open-i *(https://openi.nlm.nih.gov)*, which was accessed on December 11, 2023 (20). Subsequently, reports were excluded due to missing demographic information and missing or blank report files (Fig 1A). This data set is an ideal clinical LLM exemplar use-case as it contains 13 well-defined standardized radiologist-labeled reference standard disease labels, allowing for structured evaluation of LLM accuracy in clinical language interpretation, such as accurately extracting the presence of a specific finding or disease (21). Further details on how this data set was generated have been described previously (20).

Instruction Prompt

The 13 CheXpert (Stanford ML Group; *https://stanfordmlgroup.github.io/competitions/chexpert/*) labels were used to develop a single-task prompt (Appendix S1), which was essentially an adaptation of a prior published prompt (9), explaining CheXpert clinical terminology and instructing each LLM to confirm the presence or absence of each of the 13 labels across the entire data set (Fig 1B). Two expert clinical radiologists verified the accuracy of the CheXpert terminology. Of note, the labels are not independent of each other, and can overlap as they are not mutually exclusive. The following are the 13 CheXpert labels used within the prompt in alphabetical order: atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, and support devices. The "no finding" label was excluded, as it is derived from the positive or uncertain presence of the other labels. Within the prompt, the model was instructed to categorize the 13 possible findings as positive (value of 1) or negative (value of 0) in JavaScript Object Notation standard text-based format (Fig 2).

For the given task and prompt, each model was tested on whether it was able to produce valid JavaScript Object Notation (JSON) files with no missing labels (adequate response) or failed to follow instructions with missing labels in the JSON file (inadequate response). Individual label accuracy and combined overall accuracy for the report were calculated, given that there was potential for overlapping labels.

Energy Use Evaluation Experiments

Each model generated a response on the entire data set (overall accuracy), and subsequently individually identified the presence or absence of 13 disease labels. CodeCarbon (version 2.3.1; BCG GAMMA, Comet.ml, Haverford College, MILA), an open-source software tool designed for tracking the energy use (in kilowatt-hours) associated with computing tasks, was used to track the energy use of each LLM during the experiments (22–24). CodeCarbon records the energy use of the computing environment inclusive of graphics processing units (GPUs), central processing units, and random access memory. However, the experiments in this study focused solely on GPU energy consumption as many AI tools, including LLMs, rely heavily on GPUs for their accelerated parallel processing tasks. This approach allowed for more streamlined data collection and more precise evaluation of LLM energy use to better inform future AI optimization efforts by concentrating on the primary energy-consuming component in AI workloads. GPU energy use was measured using the pynvml library that was automatically installed with CodeCarbon (23).

Performance accuracy and associated energy use on the entire data set was obtained across five parameter sizes between the two LLM models (Llama 2 at 7B, 13B, and 70B parameters; Vicuna 1.5 at 7B and 13B parameters), which were run on the default settings with the single-task instruction prompt for the 13 CheXpert labels described above. All models were run on local compute clusters with NVIDIA RTX A6000 (NVIDIA; *https://www.nvidia.com/en-us/design-visualization/rtx-a6000/*) GPUs. While a single GPU was used for almost all the models, the larger Llama 2 70B model required four NVIDIA RTX A6000 GPUs to run due to computational demands.

Statistical Analysis

Accuracy was calculated for each CheXpert label separately and complemented by sensitivity, specificity, and F1 metrics for each label. Overall accuracy was reported as the mean of individual accuracies of all 13 labels, reflecting a comprehensive perspective on model performance across all diagnostic labels. The tradeoff in efficiency of this performance was defined and calculated as an "efficiency ratio" (performance per energy unit, or accuracy per kilowatt-hour, where accuracy was the overall accuracy) for each model type and size. For the efficiency ratio, overall accuracy was chosen for the use-case as it measures the model's ability to handle multiple labeling tasks simultaneously, providing a more holistic view of performance-energy tradeoffs for the global task. Data were initially generated in Python 3.12.4 (Python Software Foundation), and then descriptive statistics were generated in Excel (version 2405, build 16.0.17628.20006, 64-bit; Microsoft 365) for analysis and visual comparison.

Reference Standard Report

COMPARISON: PA and lateral chest
INDICATION: XXXX-year-old female with breast mass and smoking history.
FINDINGS: The heart size and cardiomediastinal silhouette are normal. There is hyperexpansion of the lungs with flattening of the hemidiaphragms. There is no focal airspace opacity, pleural effusion, or pneumothorax. There multilevel degenerative changes of thoracic spine.
IMPRESSION: Emphysema, however no acute cardiopulmonary finding.

	Individual Labels													
Model (billion parameters)	Enlarged cardiom	Cardiomegaly	Lung Lesion	Lung Opacity	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural Other	Fracture	Support Devices	Accuracy
Reference Standard Label	0	0	1	0	0	0	0	0	0	0	0	0	0	
LLaMa-2 (7B)	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	Inadequate
LLaMa-2 (13B)	0	1	0	0	0	0	0	0	0	0	0	0	0	84.62% (11/13 labels)
LLaMa-2 (70B)	0	0	0	0	0	0	0	0	0	0	0	0	0	92.31% (12/13)
Vicuna-1.5 (7B)	0	0	0	0	0	0	0	0	0	0	0	0	0	92.31% (12/13)
Vicuna-1.5 (13B)	0	0	0	0	0	0	0	0	0	0	0	0	0	92.31% (12/13)

Figure 2: Illustration shows an example output performed on a single-prompt task of identifying 13 labels on chest radiograph reports. This output shows the performance for individual and global labeling tasks by the Meta's Llama 2 (7B, 13B, 70B) and LMSYS Org's Vicuna 1.5 (7B, 13B) models compared with the 13 CheXpert radiologist-labeled reference standard disease labels. Within the prompt (Appendix S1), the model is instructed to categorize the 13 possible findings as positive (value of 1) or negative (value of 0) in a JavaScript Object Notation standard text-based format. The reference standard labels (0) are shown in black, with inadequate responses (−1) in grey. An inadequate response indicates the model failed to follow instructions and had missing labels. The large language model outputs that are concordant with reference standard findings are shown in green, and those that are discordant are in orange. B = billion, Cardiom = cardiomediastinum, PA = posteroanterior.

Results

Characteristics of the Data Set

Of 3955 chest radiograph reports, 290 were excluded due to missing demographic information (n = 267) and missing or blank report files (n = 23), leaving 3665 chest radiograph reports evaluated in the study (Fig 1A).

Accuracy and Valid Outputs

The diagnostic performance of each model and parameter size according to the 13 disease labels is illustrated in Figure 3, and additional individual-label results for Llama 2 and Vicuna 1.5 are provided in Tables 1 and 2, respectively. Llama 2 demonstrated increased accuracy with increasing parameter size as follows: the 7B model achieved an overall accuracy of 7.91% (289.76 of 3665 reports), with inadequate responses at 20.22% (741 of 3665 reports); the 13B achieved an overall accuracy of 74.08% (2715.15 of 3665 reports), with inadequate responses at 5.38% (197 of 3665 reports); and the 70B model reached an overall accuracy of 92.70% (3397.38 of 3665 reports), with inadequate responses at 0.03% (one of 3665 reports) (Table 3). The fine-tuned Vicuna 1.5 model showed consistently high accuracy levels, with the 7B and 13B sizes achieving overall accuracies of 93.83% (3438.69 of 3665 reports) and 93.65% (3432.38 of 3665 reports), respectively, and a similar proportion of inadequate responses at 0.03% (one of 3665 reports).

Energy Use

The Llama 2 models consumed more energy, with the 7B, 13B, and 70B models using 0.59 kWh, 1.81 kWh, and 4.16 kWh, respectively. In contrast, the Vicuna 1.5 models used less energy overall (7B, 0.13 kWh; 13B, 0.28 kWh) (Table 3).

Efficiency Ratios

The present study introduces an efficiency ratio (accuracy per kilowatt-hour) as a novel metric to evaluate the practical utility of LLMs against their energy consumption. The Vicuna 1.5 models emerged as the most efficient, with 7B achieving the highest ratio (737.28) and the 13B model close behind (331.40), while all Llama 2 models had much lower efficiency ratios (7B, 13.39; 13B, 40.90; 70B, 22.30) (Table 3).

Discussion

Large language models (LLMs) show inherent tradeoffs between accuracy and energy use in an exemplar medical application, highlighting the need to select the right model for the right task to balance performance and resource usage. Thus, we evaluated five different parameter sizes of two open-source LLMs using chest radiograph reports from the Indiana University Chest X-ray Collection available from the National Institutes of Health–National Library of Medicine Open-i search engine. Using data from 3665 chest radiograph reports, we established an efficiency ratio (accuracy per kilowatt-hour) to evaluate the practical utility of LLMS against energy consumption.

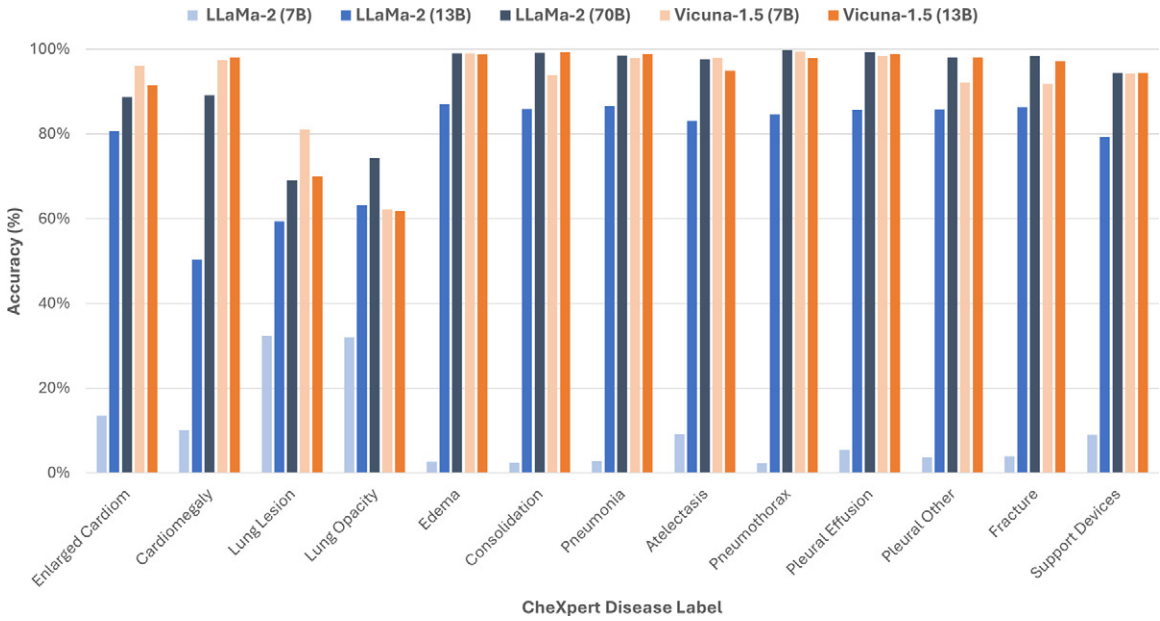

Figure 3: Bar graph shows the label-wise comparative diagnostic accuracy of each large language model (LLM) parameter size according to 13 CheXpert disease labels (the reference standard), illustrating the performance of general-purpose (Llama 2; Meta) and fine-tuned (Vicuna 1.5; LMSYS Org) LLMs in the same prompted task. B = billion, Cardiom = cardiomediastinum.

Two models were evaluated as follows: Meta's Llama 2 (16), a general-purpose model, and LMSYS Org's Vicuna 1.5 (17), a specialized fine-tuned model. The Vicuna 1.5 7-billion (B)–parameter model size had the highest efficiency ratio (737.28 vs 13.39 for Llama 2 7B), while the Vicuna 1.5 7B and 13B models had the highest overall diagnostic accuracy at 93.83% and 93.65%, respectively. The larger Llama 2 70B model used more than seven times the energy of its 7B counterpart (4.16 kWh vs 0.59 kWh)with lower overall accuracy, resulting in a low efficiency ratio of 22.30. This challenges the assumption that larger, more complex models will have superior performance, and instead demonstrates that fine-tuned models can achieve high accuracy with lower energy consumption. Overall, these measurements help set an efficiency benchmark for LLM inference tasks in energy-conscious clinical settings. Our energy use findings are comparable with that of existing literature; for example, assessment of another LLM showed that an inference task using the Multi-Genre Natural Language Inference (MultiNLI) corpus consumed approximately 3.2 kWh of electricity, which is within range of our measurements (25).

However, between model sizes, this accuracy versus energy use tradeoff is not straightforward. For example, the Vicuna 7B model unexpectedly outperformed its 13B counterpart in accuracy, contrary to typical expectations that larger models yield better performance. We hypothesized that the specific nature of our medical labeling task—focusing on instruction-based processing—might align more closely with the optimization parameters of the 7B model. This suggests that, beyond size, the specific architecture and training of a model play critical roles in its suitability for particular medical tasks. Such insights prompt a deeper examination of how fine-tuning and model complexity interact with task-specific requirements. Additionally, the Vicuna 13B model had higher energy use compared with the Vicuna 7B model, and overall may be a poorer choice for both performance and sustainability if just comparing these two models. This underscores the importance of considering both model architecture and task specificity when selecting LLMs for clinical applications.

Generative AI tools (eg, LLMs) have increased AI accessibility and applications in health care, simultaneously raising new concerns of equitable distribution of computational resources and their downstream consequences, including energy costs and availability (22,26). The insights from the present study contribute valuable data to this ongoing discussion around the sustainability of AI in radiology, showing the nuanced relationship between LLM type, model size, accuracy, and energy use. When considering LLMs for clinical applications, there should be balance between achieving high diagnostic accuracy and energy sustainability, especially at large AI adoption scales. By carefully considering model characteristics, such as type and size, alongside computational demands, researchers and clinicians can make informed environmentally friendly choices.

We acknowledge that AI accuracy and other clinical performance metrics will most certainly be prioritized in medical settings to ensure patient safety, and tradeoffs in accuracy for gains in energy use may not be practical or realistic. However, it is important to recognize that the pursuit of maximizing AI accuracy should not occur in isolation from considerations of energy use, and ideally there should be attention and effort towards improving both accuracy and energy use. In research and clinical settings, performance reporting should include both accuracy and energy consumption metrics for a comprehensive understanding of an AI model's sustainability impact. This is especially relevant in high-volume AI screening applications or for continuous monitoring systems that operate around the clock, where cumulative energy savings may be

Table 1: Individual-label Results Used in Accuracy Calculations for Meta's Llama 2 across Three Model Parameter Sizes for 13 Common Disease Labels

Disease and Model Parameter Size	TP	FP	TN	FN	Sensitivity (%)	Specificity (%)	F1 Score (%)	Accuracy (%)
Atelectasis								
7B	225	2653	42	5	97.83	1.56	16.70	9.13
13B	151	455	2732	131	53.55	85.72	65.13	83.11
70B	244	35	3334	52	82.43	98.96	89.39	97.63
Cardiomegaly								
7B	262	2630	32	1	99.62	1.20	18.26	10.05
13B	226	1655	1519	69	76.61	47.86	60.73	50.30
70B	322	395	2944	4	98.77	88.17	93.69	89.11
Consolidation								
7B	23	2855	47	0	100.00	1.62	4.67	2.39
13B	24	482	2963	8	75.00	86.01	80.08	85.91
70B	25	30	3609	1	96.15	99.18	97.63	99.15
Edema								
7B	31	2848	45	1	96.88	1.56	5.06	2.60
13B	27	439	2992	11	71.05	87.20	78.23	87.03
70B	41	33	3588	3	93.18	99.09	96.01	99.02
Enlarged cardiomediastinum								
7B	302	2522	92	9	97.11	3.52	23.66	13.47
13B	197	519	2601	152	56.45	83.37	66.41	80.66
70B	341	376	2909	39	89.74	88.55	89.20	88.68
Fracture								
7B	65	2811	49	0	100.00	1.71	7.50	3.90
13B	37	432	2957	43	46.25	87.25	60.23	86.31
70B	82	55	3526	2	97.62	98.46	98.03	98.44
Lung lesion								
7B	919	1968	28	10	98.92	1.40	48.79	32.38
13B	107	410	1952	1000	9.67	82.64	16.62	59.35
70B	40	0	2491	1134	3.41	100.00	6.49	69.06
Lung opacity								
7B	904	1979	32	10	98.91	1.59	48.36	32.00
13B	360	554	1834	721	33.30	76.80	43.63	63.25
70B	215	6	2508	936	18.68	99.76	29.85	74.30
Pleural effusion								
7B	110	2765	47	3	97.35	1.67	10.17	5.37
13B	75	437	2898	59	55.97	86.90	67.72	85.70
70B	131	13	3507	14	90.34	99.63	94.59	99.26
Pleural other								
7B	56	2819	49	1	98.25	1.71	6.93	3.59
13B	1	426	2976	66	1.49	87.48	2.93	85.82
70B	2	1	3592	70	2.78	99.97	5.40	98.06
Pneumonia								
7B	32	2845	48	0	100.00	1.66	5.32	2.74
13B	9	443	2994	23	28.13	87.11	42.46	86.57
70B	33	53	3576	3	91.67	98.54	94.95	98.47
Pneumothorax								
7B	17	2860	48	0	100.00	1.65	4.35	2.22
13B	10	521	2925	13	43.48	84.88	57.44	84.61
70B	19	0	3640	6	76.00	100.00	86.30	99.84
Support devices								
7B	218	2657	44	6	97.32	1.63	16.40	8.96
13B	80	535	2670	184	30.30	83.31	43.85	79.27
70B	93	13	3366	193	32.52	99.62	48.37	94.38

Note.—Except where indicated, data are numbers of chest radiograph reports. A total of 3665 reports were obtained from the Indiana University Chest X-ray Collection available from the National Library of Medicine's Open-i. The 13 diseases are derived from CheXpert radiologist-labeled disease labels as the reference standard. B = billion, FN = false negative, FP = false positive, TN = true negative, TP = true positive.

Table 2: Individual-label Results Used in Accuracy Calculations for LMSYS Org's Vicuna 1.5 across Two Model Parameter Sizes for 13 Common Disease Labels

Disease and Model Parameter Size	TP	FP	TN	FN	Sensitivity (%)	Specificity (%)	F1 Score (%)	Accuracy (%)
Atelectasis								
7B	280	66	3303	8	97.22	98.04	97.60	97.98
13B	289	178	3191	7	97.64	94.72	96.28	94.95
Cardiomegaly								
7B	314	83	3256	12	96.32	97.51	96.86	97.41
13B	260	5	3334	66	79.75	99.85	87.97	98.06
Consolidation								
7B	16	214	3424	10	61.54	94.12	74.35	93.89
13B	26	25	3614	0	100.00	99.31	99.66	99.32
Edema								
7B	38	30	3591	6	86.36	99.17	92.26	99.02
13B	43	43	3578	1	97.73	98.81	98.26	98.80
Enlarged cardiomediastinum								
7B	305	69	3216	75	80.26	97.90	87.46	96.07
13B	66	0	3285	314	17.37	100.00	29.19	91.43
Fracture								
7B	78	299	3281	0	100.00	91.65	95.74	91.83
13B	82	102	3479	2	97.62	97.15	97.39	97.16
Lung lesion								
7B	538	63	2428	634	45.90	97.47	58.59	80.97
13B	76	0	2491	1098	6.47	100.00	11.85	70.04
Lung opacity								
7B	552	82	552	590	48.34	87.07	54.38	62.16
13B	592	172	592	559	51.43	77.49	56.15	61.83
Pleural effusion								
7B	118	41	3479	17	87.41	98.84	92.58	98.41
13B	137	35	3485	8	94.48	99.01	96.61	98.83
Pleural other								
7B	17	232	3361	55	23.61	93.54	37.59	92.17
13B	1	0	3593	71	1.39	100.00	2.74	98.06
Pneumonia								
7B	35	76	3553	1	97.22	97.91	97.56	97.90
13B	33	39	3590	3	91.67	98.93	95.12	98.85
Pneumothorax								
7B	21	15	3625	4	84.00	99.59	91.09	99.48
13B	23	75	3565	2	92.00	97.94	94.86	97.90
Support devices								
7B	119	43	3336	167	41.61	98.73	57.73	94.27
13B	93	13	3366	193	32.52	99.62	48.37	94.38

Note.—Except where indicated, data are numbers of chest radiograph reports. A total of 3665 reports were obtained from the Indiana University Chest X-ray Collection available from the National Library of Medicine's Open-i. The 13 diseases are derived from CheXpert radiologist-labeled disease labels as the reference standard. B = billion, FN = false negative, FP = false positive, TN = true negative, TP = true positive.

substantial. Additionally, lower-energy AI models may be beneficial or preferred in resource-limited settings where energy availability constrains technology use, or in more portable devices that require less energy consumption. Because accuracy is commonly reported while energy use is frequently overlooked, adopting standardized reporting of both is crucial and should become the norm (27,28). In an ideal world, clinicians and health care leaders would be presented with choices between models that do not compromise on accuracy but may vary in their sustainability. Informed decision-making should eventually include selection of an AI model based on both its clinical efficacy and its environmental footprint.

Our study had limitations. First, despite a standardized approach to terminology (CheXpert) and expert clinical radiologist discussion, developing an input prompt has the potential for subjectivity. Second, to standardize comparability across models, we used the base models that had inherent limitations (ie, context length) and did not include additional

Table 3: Overall Comparison of Meta's Llama 2 and LMSYS Org's Vicuna 1.5 across Five Parameter Sizes When Applied to the Same Chest Radiographic Data Set with the Same Prompt for Disease Label Identification

Variable	Llama 2*			Vicuna 1.5†	
	7 Billion Parameters	13 Billion Parameters	70 Billion Parameters	7 Billion Parameters	13 Billion Parameters
Inadequate response‡§	20.22 [741/3665]	5.38 [197/3665]	0.03 [1/3665]	0.03 [1/3665]	0.03 [1/3665]
Overall accuracy‡‖	7.91 [289.76/3665]	74.08 [2715.15/3665]	92.70 [3397.38/3665]	93.83 [3438.69/3665]	93.65 [3432.38/3665]
Required no. of GPUs	1	1	4	1	1
Duration (sec)	7433	22402	15787	1584	3446
GPU energy consumed (kWh)	0.59	1.81	4.16	0.13	0.28
Efficiency ratio (accuracy per kWh)	13.39	40.90	22.30	737.28	331.40

Note.—A total of 3665 reports were obtained from the Indiana University Chest X-ray Collection available from the National Library of Medicine's Open-i. See Appendix S1 for the prompt, which used 13 CheXpert radiologist-labeled disease labels as the reference standard. GPU = graphic processing unit.

* Context length, 4000 tokens.

† Context length, 16 000 tokens.

‡ Data are percentages, with numbers of chest radiograph reports in brackets.

§ Inadequate response indicates the model failed to follow instructions and had missing labels.

‖ Overall accuracy for each model was reported as the mean of individual accuracies of all 13 labels.

accuracy-improving workarounds, such as prompt engineering with retrieval augmented generation. Finally, it was necessary to deploy four NVIDIA RTX A6000 GPUs for the Llama 2 70B model, compared with a single GPU for the other models. The need for this additional processing power highlights the increased energy demand and computational time required for larger models, and the need for careful consideration of hardware usage requirements.

In conclusion, the study results showed that smaller fine-tuned large language models (LLMs) provided a more sustainable option than large general-purpose LLMs, as they used less energy without compromising overall accuracy in a radiograph labeling task. This underscores the importance of LLM selection for medical applications. As artificial intelligence (AI) in health care continues to evolve, our energy and computational resource stewardship can help mitigate the broader downstream consequences of widespread AI use, ensuring that advancements in clinical technology align with the principles of ethical, sustainable, and responsible use.

Deputy Editor: Linda Moy
Scientific Editor: Sarah Atzen

Author contributions: Guarantors of integrity of entire study, **F.X.D., D.S., V.S.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **F.X.D., D.S., A.J., P.H.Y., V.S.P.**; clinical studies, **D.S.**; experimental studies, **F.X.D., D.S., A.K., P.H.Y., V.S.P.**; statistical analysis, **F.X.D., D.S., A.K., V.S.P.**; and manuscript editing, **F.X.D., D.S., R.C.C., A.J., P.H.Y., V.S.P.**

Disclosures of conflicts of interest: **F.X.D.** Grant funding from Association of Academic Radiology Clinical Effectiveness in Radiology Research Academic Fellowship (AAR CERRAF); honoraria for speaking at Vanderbilt University and AI for Radiology Education (AIRE). **D.S.** No relevant relationships. **A.K.** No relevant relationships. **R.C.C.** Institutional grants from the National Institutes of Health and ECOG-ACRIN; *Journal of the American College of Radiology* salary support as editor-in-chief; honoraria for grand rounds at University of Maryland, UT Southwestern, and University of Pennsylvania; travel support related to leadership and/or education roles in CERRAF, The Academy for Radiology and Biomedical Imaging Research, RSNA, and National Academies of Science, Engineering, and Medicine; data safety monitoring board for ECOG-ACRIN; board positions for AUR, the Academy, and CERRAF. **A.J.** Minimal book royalties from MIT Press. **P.H.Y.** Consultant for Bunker Hill; associate editor for *Radiology: Artificial Intelligence*; Society for Imaging Informatics (SIIM) Program Planning Committee. **V.S.P.** No relevant relationships.

References

1. Jia Z, Chen J, Xu X, et al. The importance of resource awareness in artificial intelligence for healthcare. Nat Mach Intell 2023;5(7):687–698.
2. Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI. Commun ACM 2019;63(12):54–63.
3. Dhar P. The carbon impact of artificial intelligence. Nat Mach Intell 2020;2(8):423–425.
4. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024; 30(4):1134–1142.
5. Savage CH, Park H, Kwak K, et al. General-Purpose Large Language Models Versus a Domain-Specific Natural Language Processing Tool for Label Extraction From Chest Radiograph Reports. AJR Am J Roentgenol 2024;222(4):e2330573.
6. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023;619(7969):357–362.
7. Schmidt RA, Seah JCY, Cao K, Lim L, Lim W, Yeung J. Generative Large Language Models for Detection of Speech Recognition Errors in Radiology Reports. Radiol Artif Intell 2024;6(2):e230205.
8. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. Radiology 2024;310(3):e231593.
9. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. Radiology 2023;309(1):e231147.
10. Karn SK, Ghosh R, PK, Farri O. shs-nlp at RadSum23: Domain-Adaptive Pre-training of Instruction-tuned LLMs for Radiology Report Impression Generation. In: Demner-fushman D, Ananiadou S, Cohen K, eds. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Association for Computational Linguistics, 2023; 550–556.
11. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3(1):141.

12. Patterson D, Gonzalez J, Le Q, et al. Carbon Emissions and Large Neural Network Training. arXiv 2104.10350 [preprint] https://arxiv.org/abs/2104.10350. Posted April 21, 2021. Accessed July 15, 2023.
13. Hoffmann J, Borgeaud S, Mensch A, et al. Training Compute-Optimal Large Language Models. arXiv 2203.15556 [preprint] https://arxiv.org/abs/2203.15556. Posted March 29, 2022. Accessed September 13, 2023.
14. Doo FX, Vosshenrich J, Cook TS, et al. Environmental Sustainability and AI in Radiology: A Double-Edged Sword. Radiology 2024;310(2):e232030.
15. Doo FX, Kulkarni P, Siegel EL, et al. Economic and Environmental Costs of Cloud Technologies for Medical Imaging and Radiology Artificial Intelligence. J Am Coll Radiol 2023;21(2):248–256.
16. Llama 2. Meta. https://llama.meta.com/llama2/. Accessed September 1, 2023.
17. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. LMSYS ORG. https://lmsys.org/blog/2023-03-30-vicuna. Accessed April 6, 2024.
18. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv 2307.09288 [preprint] https://arxiv.org/abs/2307.09288. Posted July 18, 2023. Accessed August 23, 2023.
19. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv 2306.05685 [preprint] https://arxiv.org/abs/2306.05685. Posted June 9, 2023. Accessed December 9, 2023.
20. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016;23(2):304–310.
21. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proc AAAI Conf Artif Intell 2019;33(01):590–597.
22. Doo FX, Parekh VS, Kanhere A, et al. Evaluation of Climate-Aware Metrics Tools for Radiology Informatics and Artificial Intelligence: Toward a Potential Radiology Ecolabel. J Am Coll Radiol 2024;21(2):239–247.
23. Courty B, Schmidt V, Goyal K, et al. mlco2/Codecarbon: V2.3.1. https://mlco2.github.io/codecarbon/. Published August 16, 2023. Accessed September 1, 2023.
24. García-Martín E, Rodrigues CF, Riley G, Grahn H. Estimation of energy consumption in machine learning. J Parallel Distrib Comput 2019;134(C):75–88.
25. Dodge J, Prewitt T, Tachet des Combes R, et al. Measuring the Carbon Intensity of AI in Cloud Instances. In: PFAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, 2022; 1877–1894.
26. Kaack LH, Donti PL, Strubell E, Kamiya G, Creutzig F, Rolnick D. Aligning artificial intelligence with climate change mitigation. Nat Clim Chang 2022;12(6):518–527.
27. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP. In: Korhonen A, Traum D, Màrquez L, eds. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019; 3645–3650.
28. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Modern Deep Learning Research. Proc AAAI Conf Artif Intell 2020;34(09):13693–13696.