SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Evaluating Performance and Environmental Impact: A Comparative Study of Large Language Models

Sanaz Vahdati, MD, Mayo Clinic; Bardia Khosravi, MD, MPH, MHPE; Bradley J. Erickson, MD, PhD, CIIP, FSIIM

Introduction/Background

Leveraging artificial intelligence (AI) for diagnostic purposes has shown promise in enhancing patient care while also raising concerns about environmental sustainability. Large language models (LLMs) are increasingly impacting the medical field by automating complex tasks and facilitating a deeper understanding of vast datasets, thus revolutionizing the approach to patient care and medical research. This study focuses on the extraction of acute cervical spine fractures from radiology reports using open-source LLMs juxtaposed with an analysis of their associated carbon emissions.

Methods/Intervention

We randomly acquired radiology reports from 1000 non-contrast cervical spine CT scans conducted between January and February 2022. After prompt optimization on 110 reports, the remaining 890 served as a test dataset to assess the model's performance. The model aimed to indicate the presence or absence of an acute cervical vertebral fracture. We calculated the carbon emissions generated by running two models, Zephyr Alpha 7 Billion and LLama3 70 Billion. We applied utilizing a carbon tracker package to assess the environmental impact of LLM’s operations.

Results/Outcome

The Zephyr7B model achieved an accuracy of 94% for extracting acute cervical spine fracture, and LLAMA3 70B obtained an accuracy of 92% for this task. The sensitivity and specificity were 0.97,0.94 and 0.97,0.91 for Zephyr 7B and LLama70B, respectively. The carbon emission analysis revealed that the inference of the Zephyr model is estimated to use Energy 0.42 kWh of electricity, contributing to 0.145 Kg of CO_2eq. The LLama3 model is estimated to use Energy 1.17 kWh of electricity, contributing to 0.42 Kg of CO_2eq. (Table1)

Conclusion

In this study, we compared the LLM models’ performance and their environmental impact. We demonstrate the potential of achieving high performance using a smaller model size, which can lead to a more environmentally sustainable application of LLMs. We highlight the trade-offs between efficiency and environmental impact on the deployment of AI in medical settings.

Statement of Impact

Our findings advocate for a balanced approach to adopting AI technologies, considering their medical benefits and ecological footprints. Future work should explore optimization techniques to reduce the energy consumption of AI systems without compromising their performance, thereby aligning AI advancements with sustainable healthcare practices.

LLM Model	Accuracy	Sensitivity	Specificity	CO2eq (Kg)	Energy (kWh)	Travel by car (Km)
Zephyr Alpha 7B	94%	%97	%94	0.145	0.42	1.41
LLama3 70B	92%	0.97	0.91	0.42	1.17	3.98

Table1. Comparative evaluation of LLM model performance and environmental impact

Table1. Comparative evaluation of LLM model performance and environmental impact

Keywords

Artificial Intelligence; Large Language models; Sustainability