Executive Summary

In the past years, the increasing computational power of Large Language Models (LLMs) such as ChatGPT or Gemini, have come with a large ecological costs. Training one state-of-the-art model can produce a carbon footprint which is comparable to hundreds of flights across the United States, while the cooling of the data centers can consume the equivalent of 1.2 million peoples average annual water consumption (Jegham et al., 2025; Strubell et al., 2019). From a more holistic perspective, the fast turnover of the specialised hardware used to train these models drives e-waste, and the vast rare mineral extractions to produce them have raised human rights concerns (Li et al., 2025). These concerns underscore the importance of developing accurate tools to assess and mitigate the full carbon footprint of LLMs.

In response to these growing concerns, the concept of ”Green AI” was developed by Schwartz et al. (2019). The authors state that research in the AI community should focus on increasing model performance alongside model efficiency. Since then, the field of ”Green AI” has developed several carbon tracking tools such as Carbontracker and CodeCarbon, and post-hoc carbon estimation tools such as MLCO2 and Green Algorithms. However, these tools all focus on estimating a models operational carbon footprint during or after training, with limited capabilities of predicting the carbon footprint before any GPU-hours have been used. This restricts their usefulness for more sustainable model tuning, early-stage project planning or compliance to new regulations such as the Corporate Sustainabiilty Reporting Directive (CSRD), requiring reporting of both scope 2 and scope 3 emissions.

This research aims to address these gaps by comparing the performance of two novel approaches to modeling the carbon footprint of LLMs. The training-phase operational carbon footprint is predicted using an architecture-aware static FLOP-based approach integrated into the LLMCarbon tool and a dynamic throughput modeling approach integrated into the OpenCarbonEval tool (Faiz et al., 2024; Yu et al., 2024). The two different approaches to modeling the embodied carbon footprint of an LLM are a full-system embodied carbon footprint integrated into LLMCarbon and a GPU-only modeling approach integrated into OpenCarbonEval. By benchmarking these contrasting approaches side-by-side, we can isolate which specific modeling approach produces more accurate results, and should therefore guide future work by researchers and practitioners. To give a more comprehensive guide into how these approaches should be used, we also analyze which key input features for these models have the most influence on their performance.

To obtain the ground-truth measurements to benchmark the predictions against, we train three Transformer models on Google Cloud Platform VM instance, with NVIDIA L4 GPU's. Utilizing CodeCarbon as an integrated tracker into our code, we can measure our training-phase operational carbon footprint. Our embodied carbon footprint ground-truth is taken from the most up-to-date data published by Li et al. (2025). Our findings show that while the modeling approaches adopted by LLMCarbon slightly outperform those adopted by OpenCarbonEval for both the training-phase and embodied carbon footprint, they both exhibit high error rates. The Mean Absolute Percentage Error (MAPE) of the training-phase predictions exceeded 95%, while for the embodied carbon footprint predictions it exceeded 92%. These results indicate that both tools are not yet reliable enough to be used for compliance reporting or carbon accounting.

To fully understand the influence of the individual input features required for the operational carbon footprint estimations, we perform several feature importance analyses. Utilising a linear regression analysis, a Sobol

analysis and a permutation feature importance (PFI) analysis, we find that the hardware efficiency has the largest influence on the performance of the carbon footprint prediction models. As the current estimations of this input feature vary widely, future modeling work should focus on creating more accurate estimations methods of hardware efficiency.

Our research aims to contribute to the existing knowledge in the Green AI domain, specifically in the benchmarking of carbon footprint prediction tools. We add to this domain by offering the first side-by-side benchmarking of novel training-phase operational carbon footprint modeling approaches and embodied carbon footprint modeling approaches. Prior work had not tested whether the dynamic modeling approach outperformed the static FLOP-based approach, our results show that they do not. We also deepen the understanding of these models by offering insights into which inputs have the most influence on model performance. This follows calls in the carbon footprint estimation domain for more insight into the drivers behind the estimation differences of carbon footprint estimation tools (Bannour et al., 2021). From a managerial perspective, our findings can help to provide a baseline or guide for companies who want to incorporate predictive carbon footprint modeling into their workflows. For companies reporting under CSRD guidelines, our research clarifies what how these tools may still be insufficiently accurate for their reporting standards, and which steps can be taken, such as improving telemetry on their hardware efficiency, to create more accurate carbon footprint calculations.

Like most research, ours does not come without its limitations. The Transformer model used was trained on a relatively small dataset due to hardware constraints, which may affect the generalizability of our results to larger LLMs. We also do not look at the inference side of the operational carbon footprint of these models, while they have been shown to be a large part of the total carbon footprint of a LLM and are part of the LLMCarbon tool (Faiz et al., 2024; Wu et al., 2022). Future research could build on this work by including models trained on larger datasets in their research. They can also provide a more holistic approach to the full carbon footprint modeling of LLMs by including inference carbon footprint accuracy benchmarking.

In conclusion, while these novel approaches to training-phase operational and embodied carbon footprint modeling present important work in the future of creating more sustainable LLMs, they are not yet sufficiently accurate for use in real-world applications. However, our identification of hadrware efficiency as the dominating influence in carbon footprint prediction performance provides a clear indication for where research should be focused when aiming to create future modeling approaches. As LLMs continue to grow and become more integrated into our daily lives, it is of critical importance that carbon footprint estimations of these models becomes as robust and standardized as the model accuracy reporting.

The code used in this thesis can be found on our Github page: https://github.com/flyppens/thesis_repo