What is the environmental impact of large language models (LLMs), and how can these same models be leveraged to reduce global carbon footprints?

Introduction English

Climate change is among the most pressing challenges facing the global community today, requiring immediate and coordinated responses across all sectors of society. Digital technology, notably artificial intelligence (AI), is simultaneously recognized as both a significant contributor to greenhouse gas emissions and as a powerful tool for addressing environmental challenges. Within the field of AI, Large Language Models (LLMs)—advanced computational models trained on vast datasets to understand and generate human-like text—have risen dramatically in prominence and application, influencing countless aspects of modern life, from digital communication to industrial optimization.

The environmental impact of these models, however, remains poorly understood by the broader public and even within many sectors of the technology industry. Their training and operation demand immense computational resources, translating into considerable energy consumption and, consequently, substantial carbon emissions. Yet paradoxically, these same technologies hold potential for significant positive environmental impacts, capable of optimizing energy use, enhancing sustainability initiatives, and supporting climate change mitigation efforts.

This thesis aims to critically analyze the dual role of Large Language Models in the context of climate change. Specifically, it addresses the central question: *"What is the environmental impact of large language models, and how can these same models be leveraged to reduce global carbon footprints?"* By systematically examining both the negative impacts associated with their development and operation, as well as their capacity for positive environmental contributions, this work seeks to offer a balanced and insightful perspective on the role of LLMs in contemporary environmental challenges. Ultimately, this exploration aims to inform sustainable practices within the field of AI, highlighting pathways for developers, policymakers, and corporations to responsibly harness the power of large-scale computational models for a more sustainable future.

Introduction French

Le changement climatique est l'un des défis les plus urgents auxquels la communauté mondiale est confrontée aujourd'hui, nécessitant des réponses immédiates et coordonnées dans tous les secteurs de la société. Le numérique, notamment l'intelligence artificielle (IA), est reconnue à la fois comme un contributeur important aux émissions de gaz à effet de serre et comme un outil puissant pour relever les défis environnementaux. Dans le domaine de l'IA, les grands modèles de langage (LLM), des modèles informatiques avancés formés sur de vastes ensembles de données pour comprendre et générer des textes semblables à ceux de l'homme, ont considérablement gagné en importance et en application, influençant d'innombrables aspects de la vie moderne, de la communication numérique à l'optimisation industrielle.

L'impact environnemental de ces modèles reste cependant mal compris par le grand public et même dans de nombreux secteurs de l'industrie technologique. Leur apprentissage et leur fonctionnement nécessitent d'immenses ressources informatiques, ce qui se traduit par une consommation d'énergie considérable et, par conséquent, par des émissions de carbone importantes. Paradoxalement, ces mêmes technologies peuvent avoir des effets positifs importants sur l'environnement, en optimisant l'utilisation de l'énergie, en renforçant les initiatives de développement durable et en soutenant les efforts d'atténuation du changement climatique.

Cette thèse vise à analyser de manière critique le double rôle des grands modèles de langage dans le contexte du changement climatique. Plus précisément, elle aborde la question centrale : **"Quel est l'impact environnemental des grands modèles de langage et comment ces mêmes modèles peuvent-ils être utilisés pour réduire l'empreinte carbone mondiale ?"** En examinant systématiquement les impacts négatifs associés à leur développement et à leur fonctionnement, ainsi que leur capacité à contribuer positivement à l'environnement, ce travail cherche à offrir une perspective équilibrée et perspicace sur le rôle des modèles linguistiques à grande échelle dans les défis environnementaux contemporains. Finalement, cette thèse vise à informer les pratiques durables dans le domaine de l'IA, en mettant en évidence les voies permettant aux développeurs, aux décideurs politiques et aux entreprises d'exploiter de manière responsable la puissance des modèles de calcul à grande échelle pour un avenir plus durable.

2.1 Large Language Models (LLMs)

Definition and Core Architecture:

Large Language Models (LLMs) are a category of advanced AI language models distinguished by their immense scale and broad capabilities. They are neural network models characterized by hundreds of millions to billions of parameters, trained on massive corpora of text . Crucially, LLMs learn using a self-supervised training paradigm: they are first trained on vast amounts of unlabeled text (for example, by predicting missing or next words in sentences), which allows them to learn linguistic patterns without manual annotation . This pre-training is typically followed by task-specific fine-tuning on labeled data, a two-stage process that leverages general language knowledge and then adapts it to particular tasks . The core architecture enabling modern LLMs is the Transformer – an architecture introduced in 2017 that relies on a self-attention mechanism to model relationships between words in a sequence . Unlike earlier recurrent neural networks, the Transformer’s self-attention allows it to capture long-range context and dependencies in text more effectively and to be trained in parallel, which has been pivotal for scaling models to unprecedented sizes . In fact, the Transformer’s excellent parallelizability and capacity have made it the de facto backbone of today’s LLMs, making it feasible to build models with tens or even hundreds of billions of parameters . Prominent LLMs such as BERT, GPT and others all adopt the Transformer architecture at their core . Once trained, an LLM can generate human-like text by sequentially predicting the most probable next word, enabling it to produce coherent sentences and paragraphs in response to a given prompt.

Evolution and Key Milestones:

LLMs did not emerge in isolation but evolved from decades of progress in language modeling. Early language models in the late 20th century were mainly statistical n-gram models that predicted text based on a fixed-length context, but these had difficulty handling long-term context and complex language structure . The shift to neural network approaches in the 1980s and 90s (e.g. recurrent neural networks and

the Long Short-Term Memory (LSTM) architecture) improved the ability to model sequential data, yet even these struggled with long-range dependencies in text . A major turning point came with the introduction of the Transformer architecture by **Vaswani et al**(google scientist attention is all you need). in 2017, which overcame many limitations of RNN/LSTM models and enabled much deeper and larger networks . This breakthrough paved the way for a new generation of pre-trained language models. In 2018 Google researchers introduced BERT (Bidirectional Encoder Representations from Transformers), a large transformer-based model that demonstrated the power of pre-training on unlabeled text and fine-tuning for NLP tasks. Around the same time, OpenAI released the first Generative Pre-Trained Transformer (GPT) model. Subsequent milestones saw an explosive growth in model size and capabilities: GPT-2 (2019) contained 1.5 billion parameters and showed remarkably fluent text generation, and GPT-3 (2020) expanded to 175 billion parameters, exhibiting emergent abilities such as few-shot learning (the capacity to perform tasks with just a few examples or instructions) that were not observed in smaller models . The launch of user-friendly LLM-driven systems like ChatGPT in late 2022 further demonstrated the practical potential of LLMs and brought them into widespread public awareness . Throughout these developments, numerous other LLMs have been introduced by both academia and industry – for example, the T5 text-to-text model, XLNet, RoBERTa, and more recently large open-source models like GPT-Neo and LLaMA – collectively pushing the boundaries of language understanding and generation. This rapid technical evolution of LLMs is seen as transformative for AI at large, significantly influencing how AI systems are built and applied .

Capabilities and Applications:

One defining feature of LLMs is their general-purpose language ability. Thanks to training on diverse, Internet-scale data, LLMs acquire a broad knowledge base and the ability to perform a wide array of language tasks. They can comprehend context and generate human-like text, enabling applications such as text generation (e.g. writing essays or articles), summarization of documents, machine translation between languages, question-answering and conversational agents, and even reasoning over information given in textual form. In contrast to earlier specialized models, a single LLM can be adapted (through fine-tuning or prompting) to succeed in many different tasks, often with minimal task-specific training data – a property that has led researchers to dub them “foundation models” for AI applications. Moreover, at very large scales, LLMs have demonstrated emergent capabilities: for instance, they can perform in-context learning, meaning the model can learn to do a new task simply by being given instructions or examples in the prompt, without additional training, a behavior not seen in smaller models . This versatility makes LLMs powerful tools across numerous domains.

LLMs are already being applied across industry and academia, driving innovation in various sectors. For example, in healthcare, LLMs can assist in analyzing and summarizing medical records or literature, and even support diagnostic decision-making by interpreting clinical notes. In finance, LLMs are used to automate report generation, analyze financial news, and detect fraudulent transactions by understanding anomalous language patterns in communication . Education stands to benefit through intelligent tutoring systems and automated grading assistants that can evaluate or even personalize feedback on student writing . In the legal domain, LLMs can help in parsing and summarizing legal documents or in supporting legal research by quickly extracting relevant precedents from large text corpora. Customer service and e-commerce have embraced LLM-driven chatbots and virtual assistants to handle queries, provide product recommendations, and personalize the user experience . Even specialized fields like automotive engineering utilize LLMs for enhancing in-car voice assistants, enabling natural language navigation commands and real-time language translation for drivers.

Across these and other domains, LLMs serve as powerful engines for automation and insight, often matching or surpassing human-level performance on specialized language tasks. By leveraging their ability to understand context and generate relevant responses, organizations are using LLMs to streamline workflows, enhance customer interactions, and unlock data-driven insights in ways that were not previously possible .

Toward Environmental Impact:

While LLMs offer remarkable capabilities, their development introduces significant sustainability challenges. The performance gains from scaling up model size come at the cost of extraordinarily high computational and energy requirements. Training a single large model (with billions of parameters) demands enormous processing power and electricity. For example, a 2019 analysis found that training one large language model could emit over 626,000 pounds of CO2 (equivalent to the lifetime emissions of five cars) . More recently, training OpenAI's GPT-3 (a 175-billion-parameter model) was estimated to consume about 1,287 MWh of electricity, producing roughly 500 metric tons of CO2 emissions . Even larger modern models (such as the more advanced GPT-4) presumably require comparable or greater resources, further exacerbating the carbon footprint of training.

Deploying and using LLMs at scale (serving millions of users) adds to ongoing energy and hardware demands. In fact, inference – the process of running the trained model for user queries – can account for a substantial share of overall energy usage. Each interaction with an LLM draws on power-intensive hardware; for instance, researchers estimate that a single ChatGPT query consumes about five times more electricity than a typical web search . Meeting the needs of countless such queries (and continuously fine-tuning models) means continuous electricity usage and cooling demands in data centers. These environmental and sustainability concerns have become a growing topic of research and debate. As we transition to the next section, we will examine the environmental impact of LLMs in more detail, considering how the pursuit of ever-larger models can be balanced with responsible and sustainable AI development.

Sources: 1. Vaswani, A., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems. 2. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. 3. Brown, T., et al. (2020). Language Models are Few-Shot Learners. NeurIPS. 4. Zhao, W. X., et al. (2023). A Survey of Large Language Models. arXiv preprint arXiv:2303.18223 . 5. Khan, A., et al. (2025). Industrial Applications of Large Language Models. Scientific Reports 15, 12345 . 6. Bommasani, R., et al. (2021). On the Opportunities and Risks of Foundation Models. Stanford CRFM. 7. MIT News (2025). Explained: Generative AI's environmental impact .