The Environmental Cost of Engineering Machine Learning-Enabled Systems: A Mapping Study

Presenter:

Kouider Chadli

Co-Author(s):

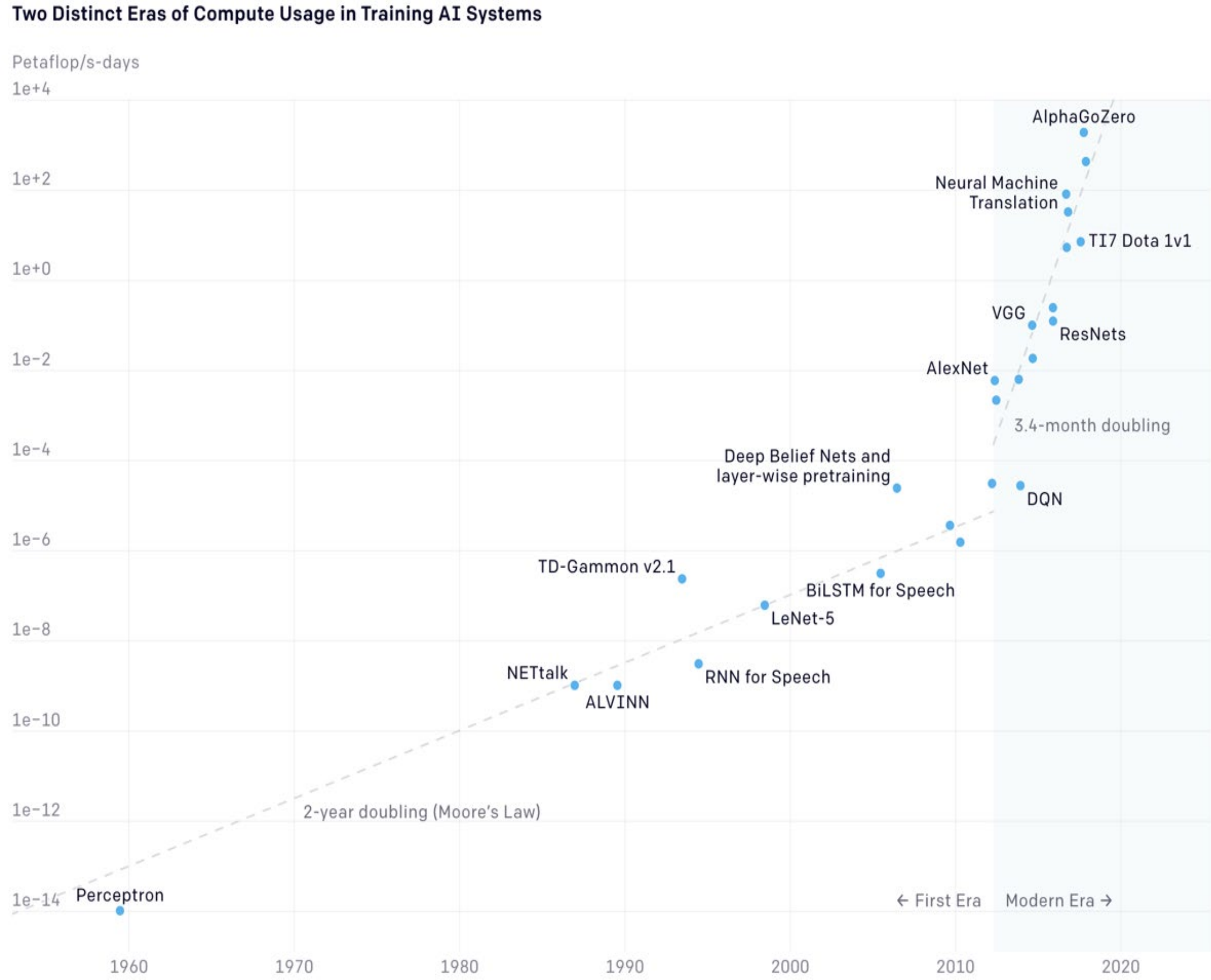
1-Dr Takfarinas Saber, University of Galway

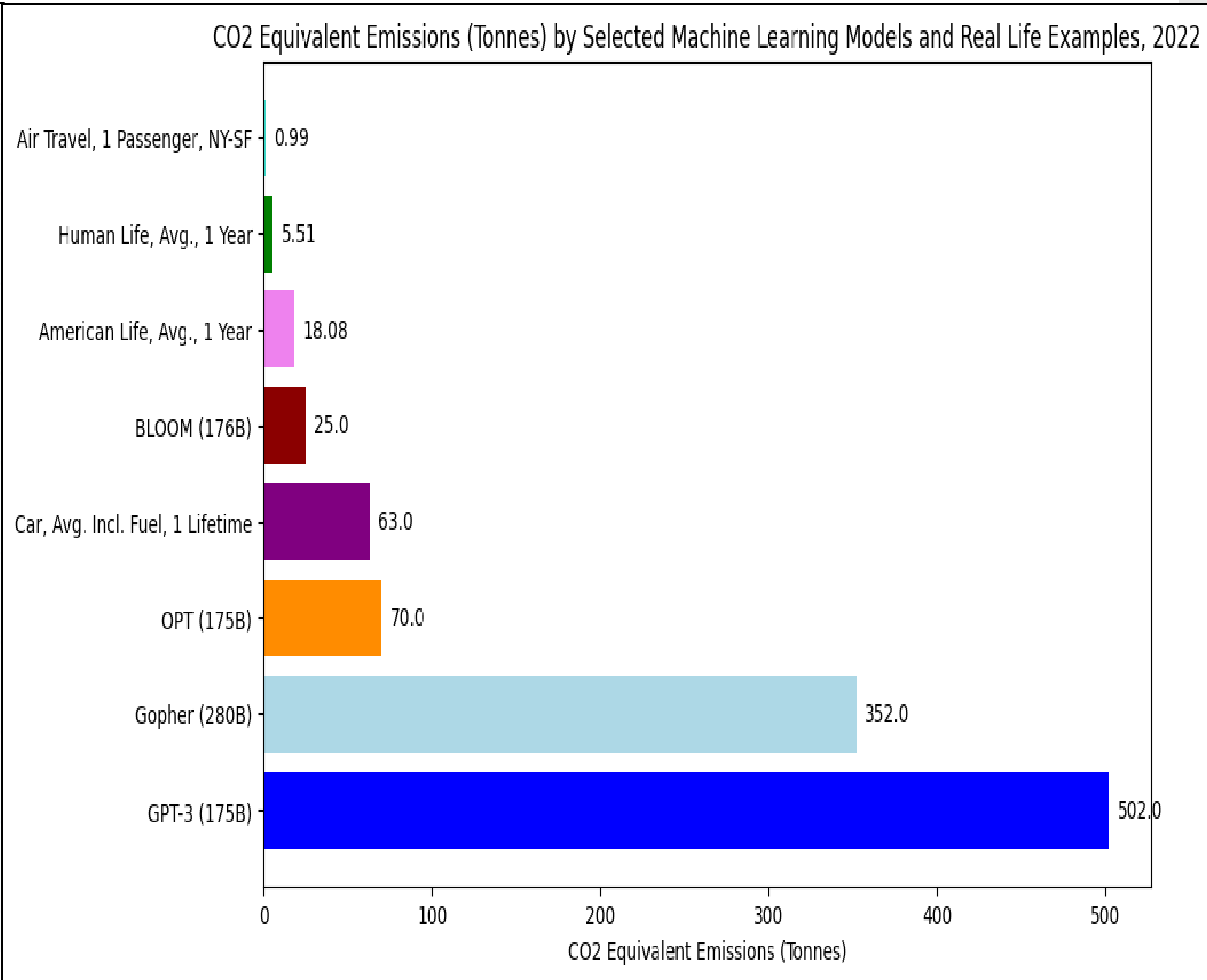
2-Prof. Goetz Botterweck, TCD

LERO Software for a better world

sfi Science Foundation Ireland

Ollscoil na Gaillimhe University of Galway





Ollscoil na Gaillimhe
University of Galway

Research Questions

- **RQ1**: What MLES domains are typically considered when assessing the environmental impact?

- **RQ2**: In which MLOps phases have environmental costs been studied?

- **RQ3**:What strategies have been proposed to assess and reduce the environmental impact associated with MLES development and operations?

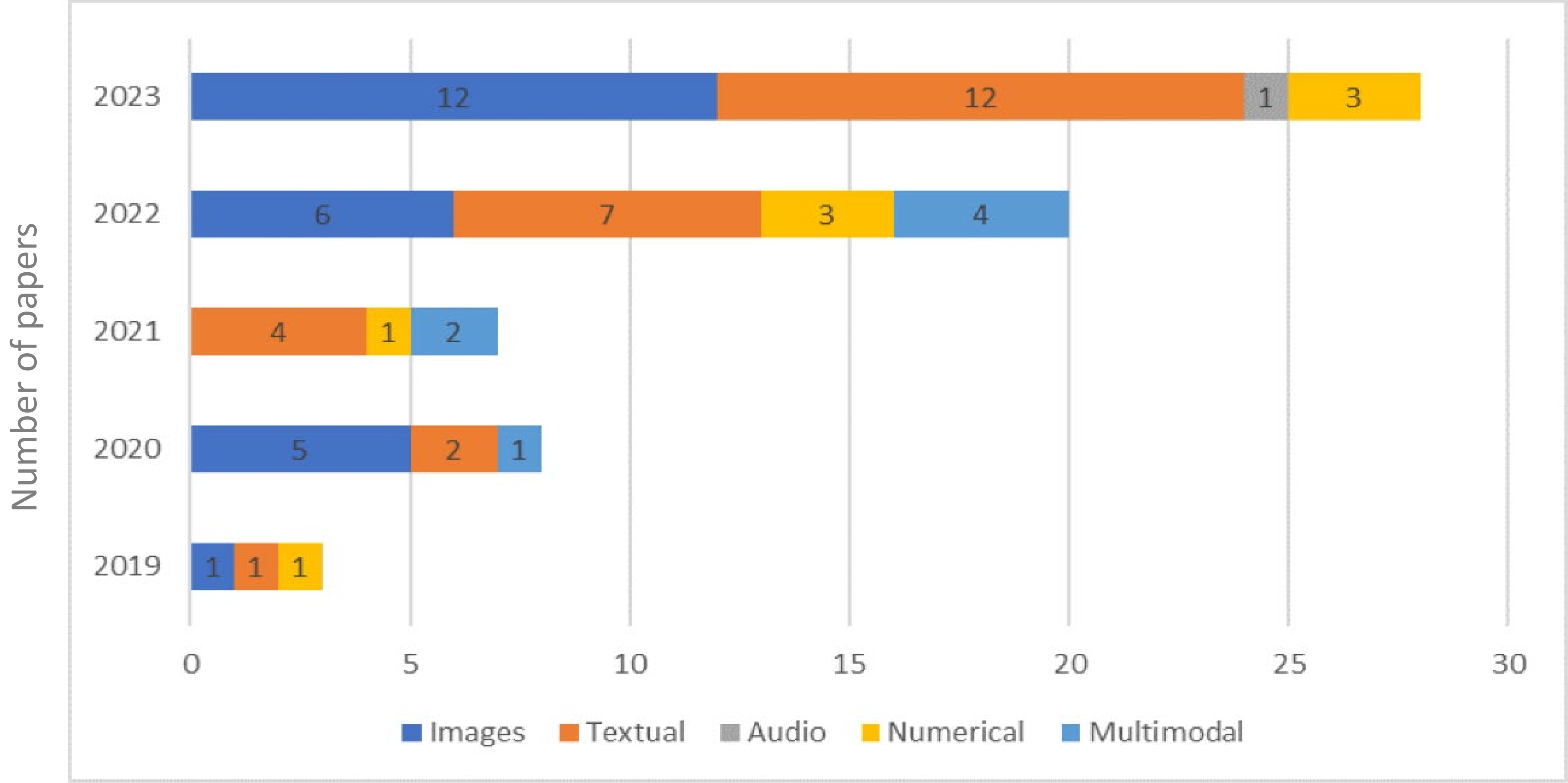
- **RQ4**:What metrics have been developed and utilized to monitor the environmental cost across MLOps?

- **RQ5**:What sustainability practices and lessons can be drawn from prior research?

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

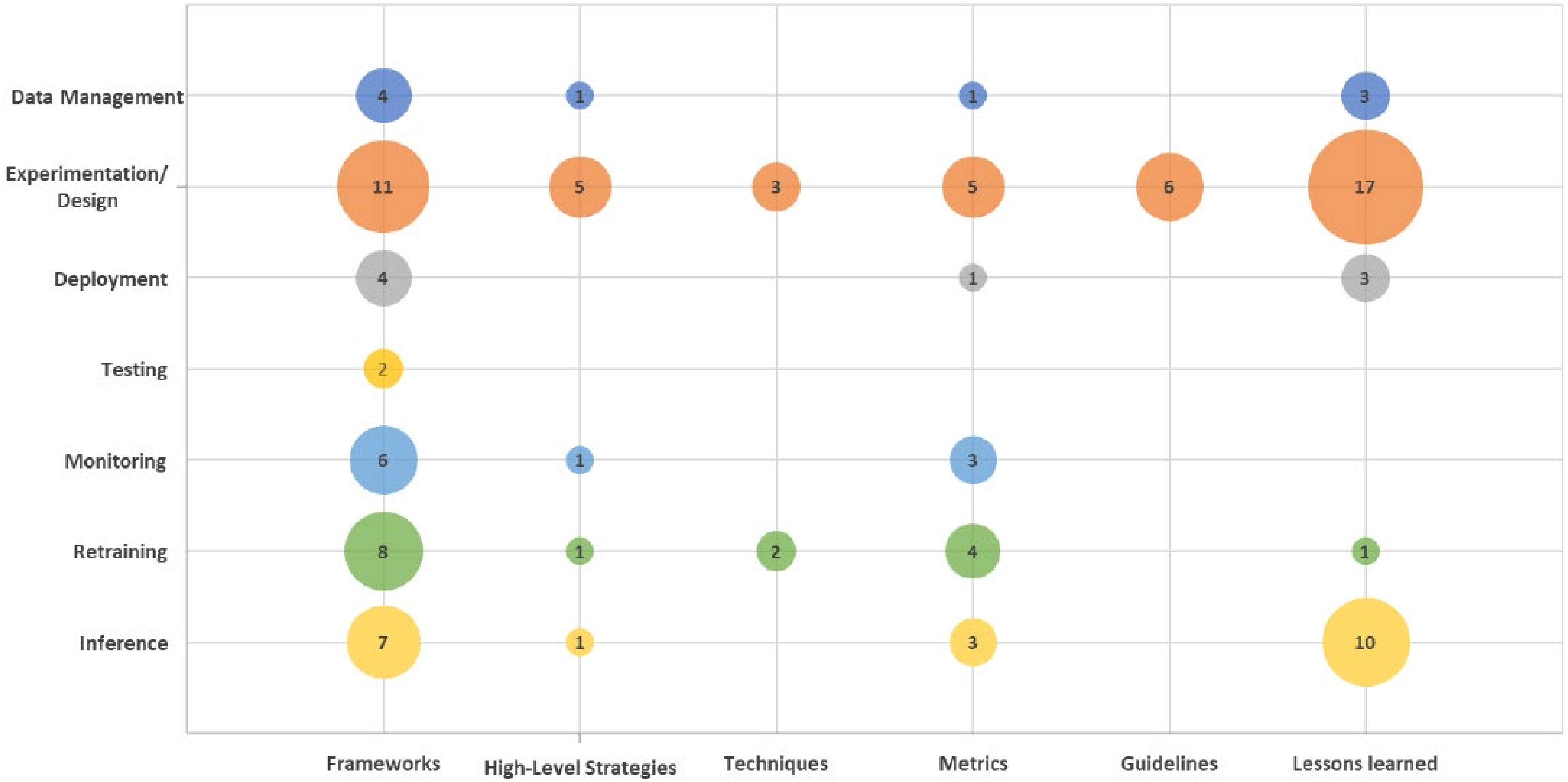
RQ1: Chronological Distribution of Identified Papers: Insights by data modality

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY



A Dual-Dimensional View: MLOps Phases vs. Contribution type

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY



Metrics to Monitor the System

MLOps Phase	Metric Description	Ref
End-to-End	An extension of the metric defined in [22].	[17]
	Introducing Software Carbon Intensity (SCI) for real-time cloud instance carbon emissions.	[16]
Experimentation and Retraining	Total power consumption estimated by combining GPU, CPU, and DRAM usage, multiplied by Power Usage Effectiveness (PUE).	[57]
Training and Inference	Deep Learning metrics gauge accuracy and energy usage.	[30]
Experimentation, Retraining and Inference	Using *CodeCarbon* [53] to estimate energy consumption by CPU and GPU.	[44]

Guidelines and lessons learned

MLOps Phase	Description of Lessons Learned and Guideline
End-to-end	Hardware energy consumption meters reveal 20% error. We need more precise measurement tools.
	There is a disparity between efficient and sustainable ML, and nuances between sustainability metrics and operational emissions. Proposed systems thinking.
	Devised guidelines to help understand the environmental implications of AI computing and mitigate its carbon footprint through optimizations in hardware, software, and operational practices.
Data Management	Reducing data input size through methods like random sampling enhances ML energy efficiency.
	Stratified sampling decreases input data features and yields energy savings.
Traininig	Recommends using random optimization over Bayesian optimization for hyperparameters as accuracy gains diminish with increased energy usage in neural network architectures.
	During multilayer perceptron classifier hyperparameter optimization, there is a point where increased energy consumption minimally improves accuracy.
	Current deep learning models are unsustainable due to their high data and computational demands. We need more efficient methods in ML to address sustainability challenges.
	Selecting energy-efficient architectures for deep learning training lowers energy usage while maintaining accuracy. The study highlights how training environments impact energy consumption and recommends factoring this in when selecting models.
	Only a minority of the 170,000 Hugging Face (HF) models report CO_2 emissions from training. Factors such as model and dataset size correlate with CO_2 emissions. Fine-tuning shows similar emissions compared to full pretraining.
	Hyperparameters in transformer models affect power consumption and model quality. Lower hidden dropout probabilities improve perplexity with minimal energy impact. While top-performing models face a trade-off between perplexity reduction and energy minimization. And increasing hidden layers increases energy usage and lowers perplexity.
	There is a link between carbon emissions, CNN architecture, and uncontrollable factors like cloud hosting location. Experimental design influences CNN training energy efficiency.
	THETA guidelines to reduce carbon emissions in model development through hyperparameter optimization, energy-efficient hardware, training logistics, and automatic mixed precision training.
	97% of the overall CO_2 emissions in Federated Learning (FL) come from client compute and client-server communications. We need energy-efficient and high-performance production FL systems.
Inference	Evaluate LLMs, examines inference performance and energy usage in distributed settings. It shows how input complexity affects performance with specific settings and hardware.
	Performance of ML models can be improved without increasing energy usage, but overall system integration/adoption may raise energy consumption (akin to better roads yielding more cars).
	GPT models have a significant environmental impact. We should prioritize sustainability in their deployment; addressing embodied carbon, and seek sustainable solutions for large model inference
	Examined energy consumption of LLMs. Identified significant variations in efficiency influenced by task, modality, model size, and architecture. stressed the trade-off between the advantages of multi-purpose systems, their energy expenditure, and resulting carbon emissions
Training and Inference	GPU energy use varies greatly. Inference with large models is energy intensive. Important to select CO_2-friendly cloud regions.
	Hardware and datacenter optimization yield substantial reduction in energy consumption for training and inference of natural language processing(NLP) apps.
	Knowledge distillation consumes 50% more energy than pre-training. Energy usage scales primarily with time and token count. In BERT-based models, inference energy costs vary with sequence lengths.
	Energy use and run-time differ for PyTorch and TensorFlow; better documentation on energy costs is needed.
	Guidelines for the ML community with established methods to estimate energy cost.
Training and Retraining	BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) training accounts for about 22% of emissions, with the rest coming from intermediate training and evaluation. Estimates of future embodied emissions will become the dominant source of emissions in ML.

Thank you

Ollscoil na Gaillimhe

University of Galway