GAINS

Green AI for Innovation and Sustainability

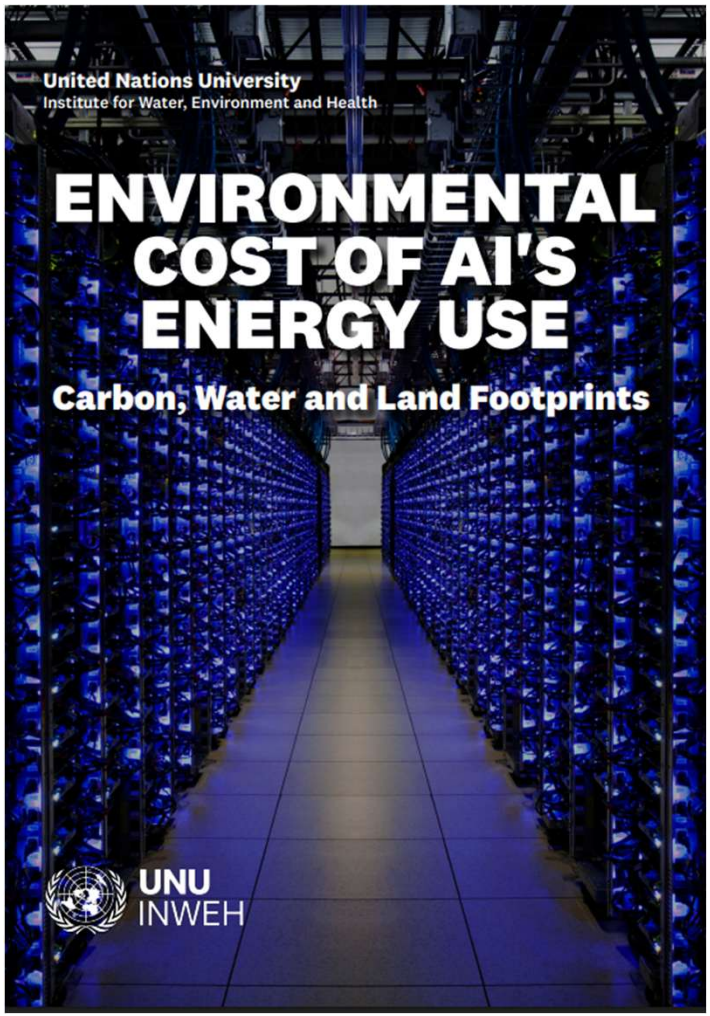
Assoc.Prof. Bernhard Geiger

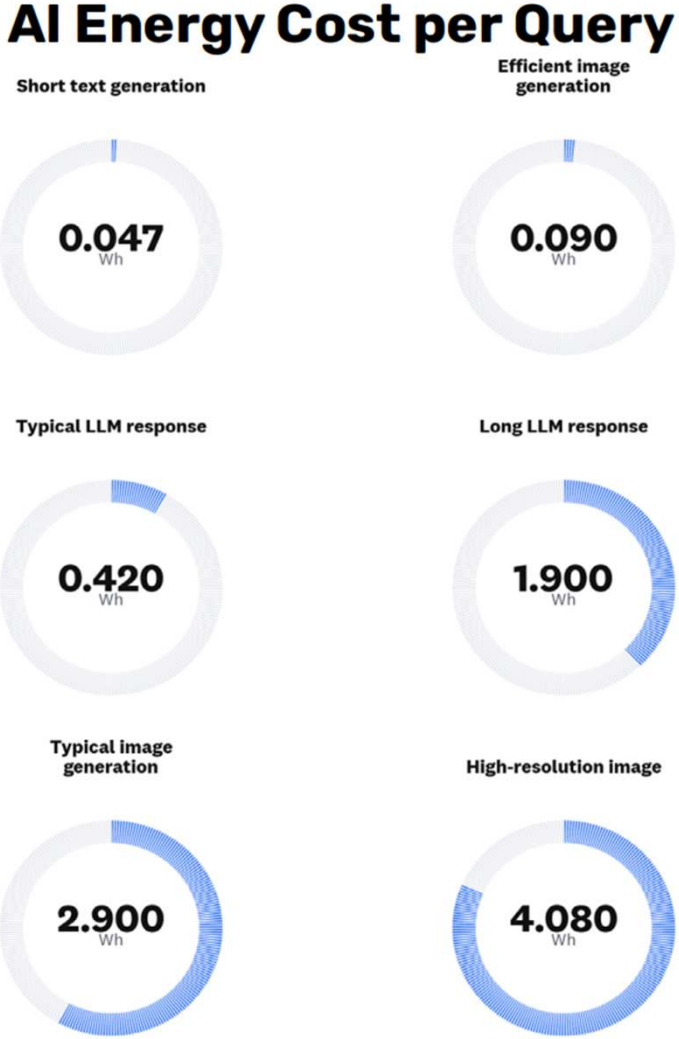
with Lolita Amann, Nicki Lisa Cole, Leonie Disch, Nikolaus Kopp, Lorenz Kutschka, Bernhard Moser, Ramin Nikzad-Langerodi, Maximilian Nothnagel, Manfred Mücke, Franz Pernkopf

Bundesministerium Innovation, Mobilität und Infrastruktur

FFG Promoting Innovation.

Dieses Projekt (Projektnummer FO999923536) wird aus Mitteln des Bundesministeriums für Innovation, Mobilität und Infrastruktur und der FFG finanziert (www.ffg.at).


United Nations University
Institute for Water, Environment and Health
ENVIRONMENTAL COST OF AI'S ENERGY USE
Carbon, Water and Land Footprints
UNU
INWEH


AI Energy Cost per Query
Short text generation
0.047
Wh
Efficient image generation
0.090
Wh
Typical LLM response
0.420
Wh
Long LLM response
1.900
Wh
Typical image generation
2.900
Wh
High-resolution image
4.080
Wh

Created with Gemini 3.5 Flash, second attempt; consumed approx. 4 Wh (50% smartphone charge) and 60 ml water.

Awareness

Key Findings from the GAINS Questionnaire

Respondent Profile (n = 42)

- Diverse stakeholder representation from **academia** (**26%**), **private companies** (**26%**), **public sector organizations** (**24%**) and **research and technology organizations** (**14%**)
- The main AI applications used in the organizations are: **LLMs (33%)**, followed by **computer vision (21%)**

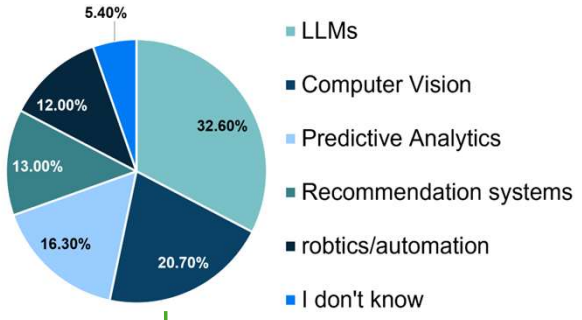
Main Findings

- **88%** are aware that AI systems require significant energy and computing resources
- **67%** report that Green AI is known at least partially within their organization
- Despite this awareness, **64%** do not currently use dedicated tools or strategies to reduce the environmental footprint of AI systems
- The main barriers/challenges to implement GreenAI practices are **limited resources and implementation effort (35%)**; , followed by **lack of knowledge (12%)**
- The strongest incentives for adopting Green AI practices are **cost savings (39%)** and **regulatory requirements (23%)**

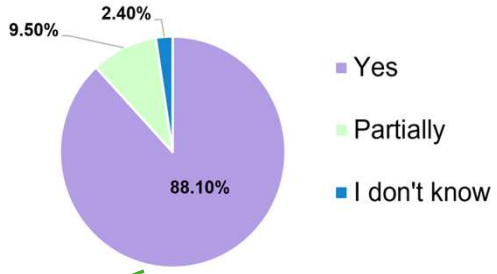
Take-away Message

Organizations are **highly aware of the environmental impact of AI**, but this **awareness has not yet translated into widespread implementation** of Green AI practices.

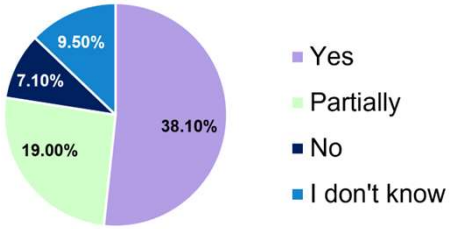
Main AI Applications in Organizations.



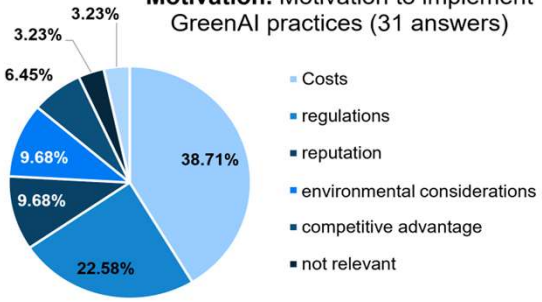
Awareness. Are Organizations aware that AI requires high energy and computing resources?



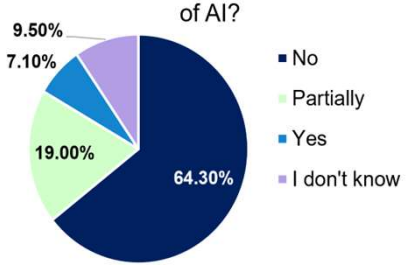
Knowledge of GreenAI. Do organizations know about GreenAI practices?



Motivation. Motivation to implement GreenAI practices (31 answers)

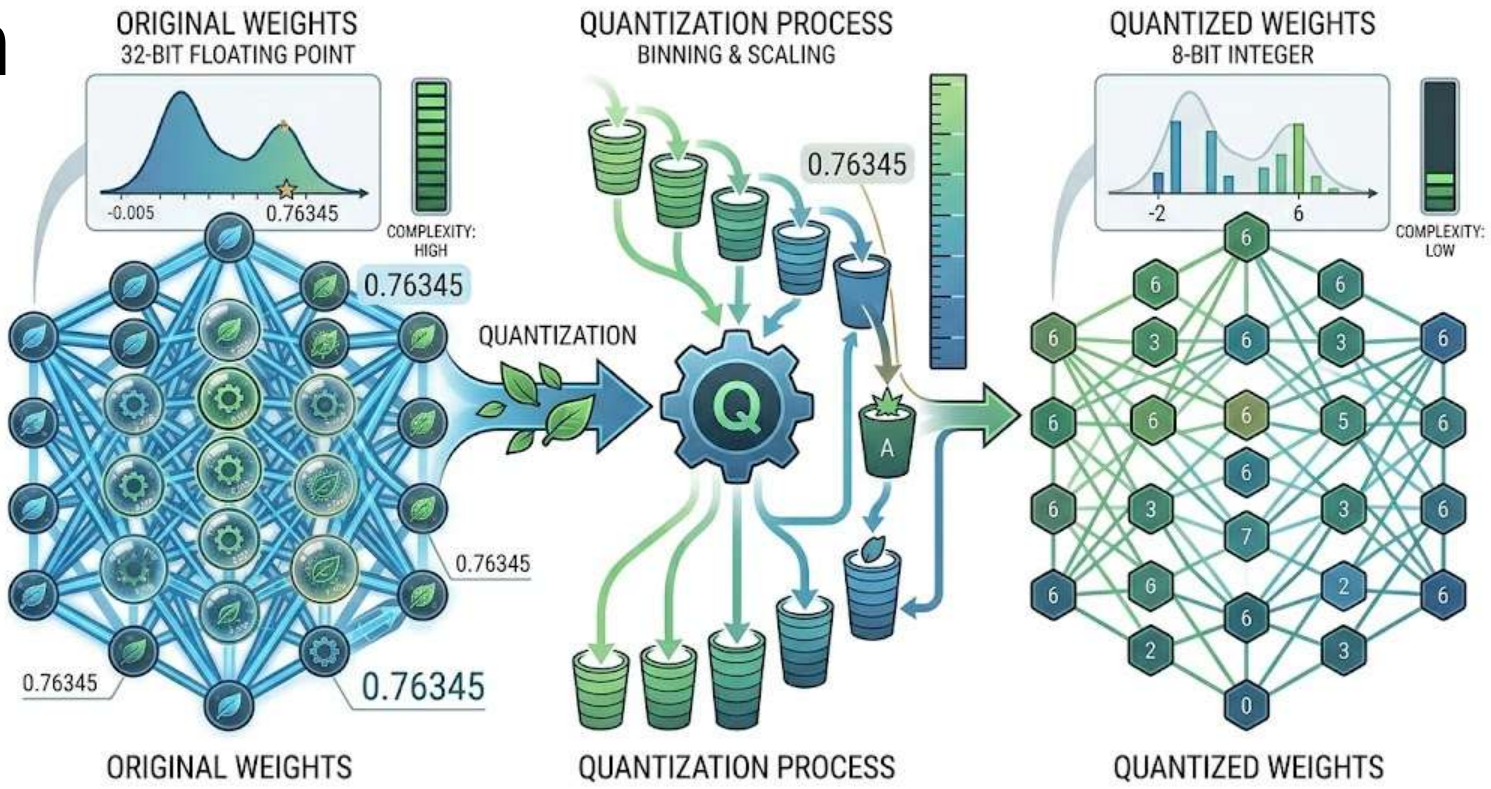


Usage of Tools. Do organizations use tools to to reduce environmental footprint of AI?



What can we do?

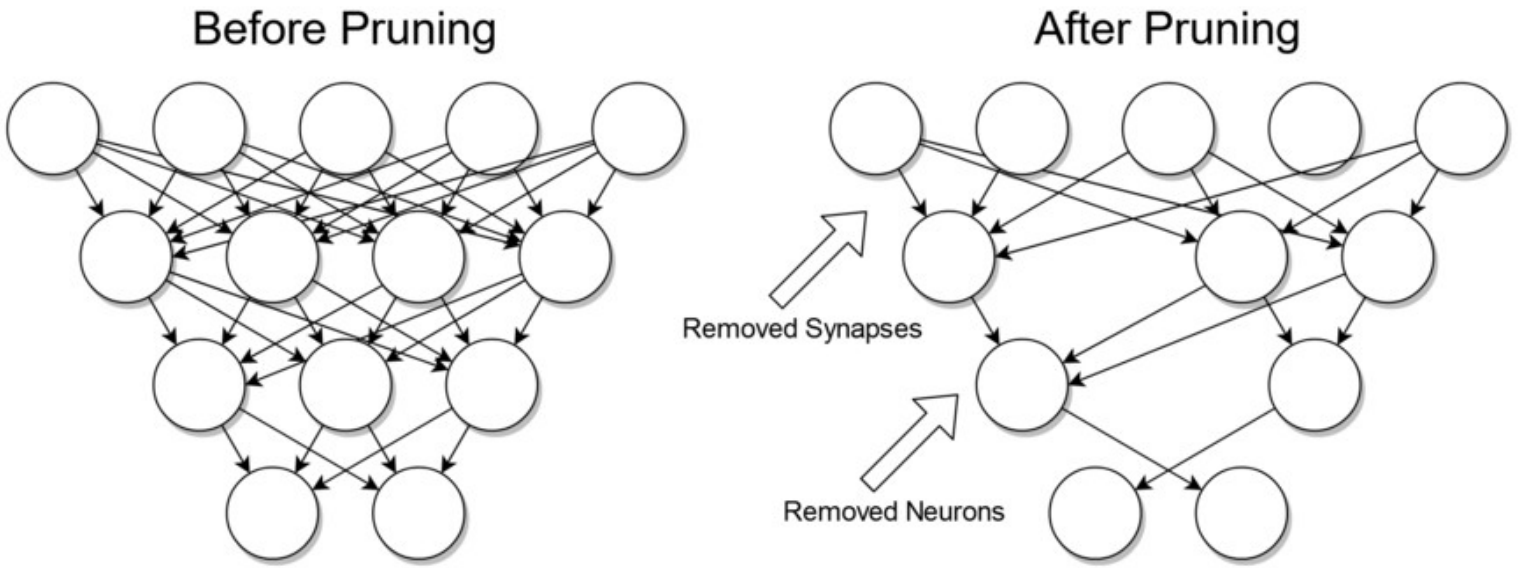
Quantization



- Replace 32/64 bit weights with 8, 4, or even 1 bits per weight
- LLM.int8() yields 8/16 bit LLMs without performance loss → up to 4-fold power reduction
- OPTQ Transformer: 100B models with 3-4 bits per weight
- 1-bit LLMs, BitNet b1.58, etc.

Created with Gemini 3.5 Flash, first attempt; approx. 2 Wh and 30 ml water

Pruning



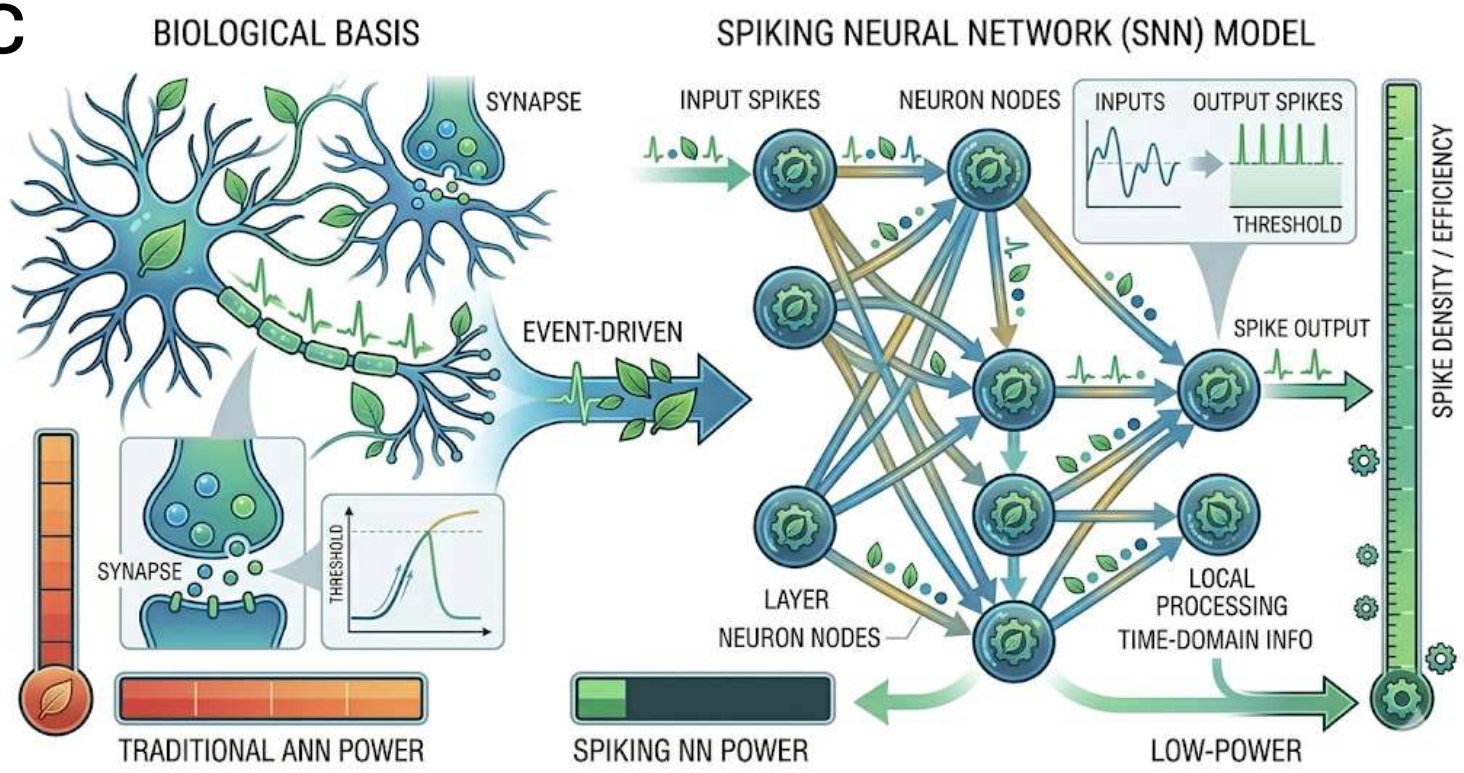
- After training (with or w/o retraining), during training, during computation
- Structured or unstructured
- Up to 90% of weights/neurons can be removed
- SparseGPT: 100B models can be reduced to half their size without performance loss
- Early-exit networks, mixture-of-experts LLMs, etc.

Neuromorphic Computing

Find out more in the upcoming **NeuroShift** report

Bundesministerium Innovation, Mobilität und Infrastruktur

FFG Promoting Innovation.



Created with Gemini 3.5 Flash, first attempt; approx. 2 Wh and 30 ml water

- Spiking Neural Networks consume energy only during spikes; extremely efficient with spike-based data sources → up to 100-fold power reduction in edge settings
- xLSTMs use efficient recurrent architectures instead of transformers
- In-memory computing performs computation in dedicated memory architectures (memristors, etc.)

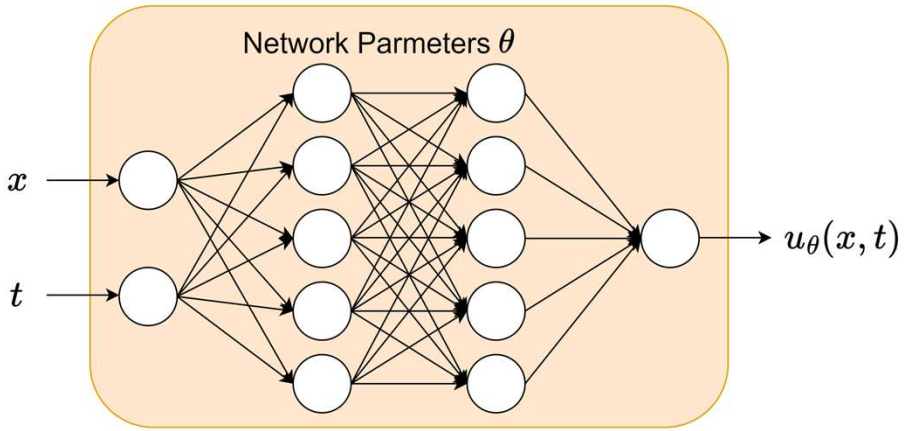
Domain-Aware Machine Learning

Differential Equation $\frac{\partial}{\partial t}u(x,t) + \mathcal{F}[u;\lambda] = 0$

Data Loss (IC/BC)

$$\mathcal{L}(\theta) = \frac{1}{|D|}\sum_{(x_i,t_i,u_i)\in D}\left(u_\theta(x_i,t_i) - u_i\right)^2 + \frac{1}{|C|}\sum_{(x_i,t_i)\in C}\left(\frac{\partial}{\partial t}u_\theta(x_i,t_i) + \mathcal{F}[u_\theta(x_i,t_i);\lambda]\right)^2$$

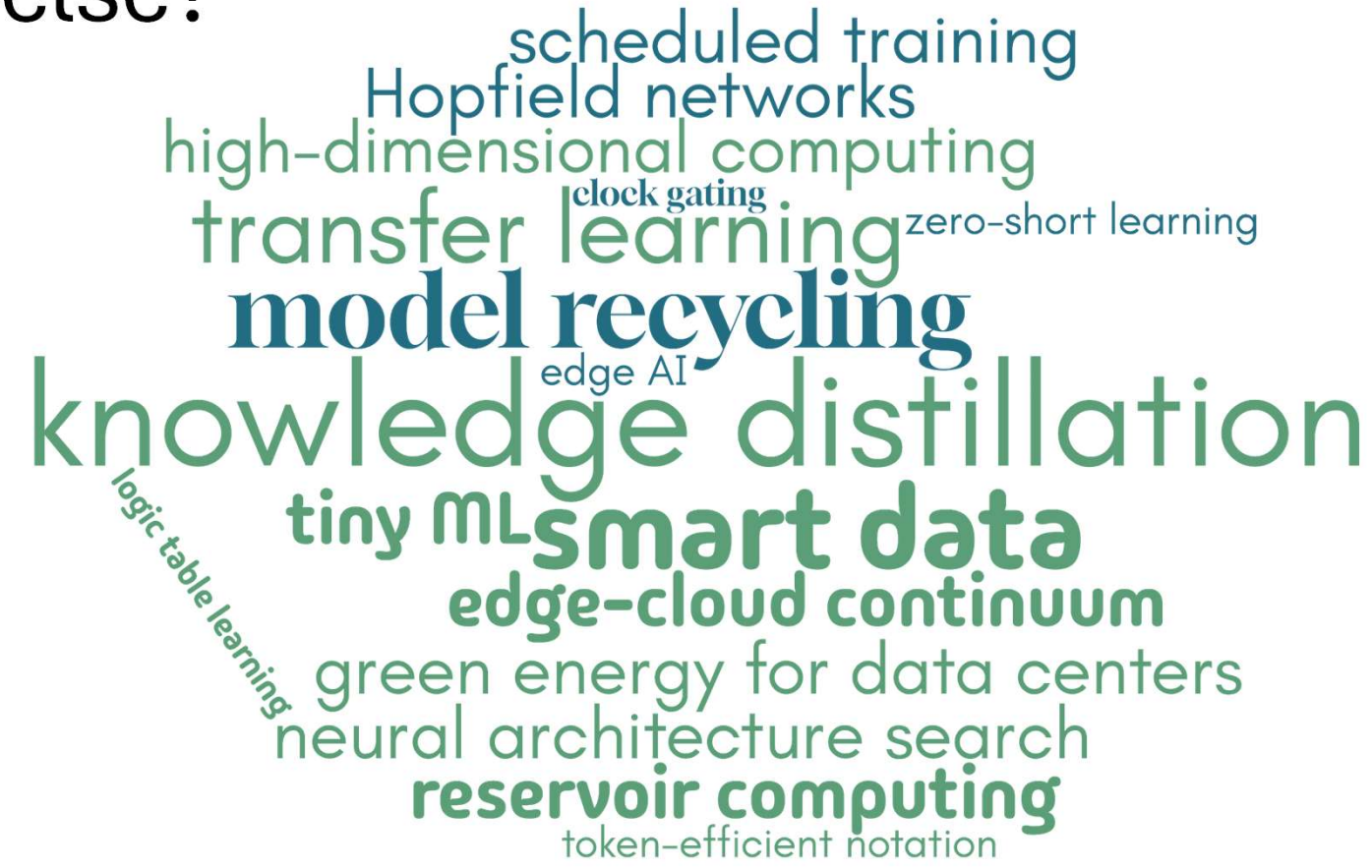
Physics Loss (consistency)



own work

- Incorporation of prior knowledge reduces data requirements and enables lower model complexity → reduced training and inference costs without performance loss
- PINNs, DeepONets, hybrid models, algorithm unrolling, etc.
- Connection to Hybrid AI

What else?



own work, https://www.wordclouds.com/

Recommendations

- **Create efficient models**: Make the models as small as possible and as large as necessary to solve the task
- **Use models efficiently**: e.g., think before you prompt, select sustainable cloud providers, don’t thank your LLM
- **Know your footprint**: Quantify and track the energy consumption of your AI models/hardware during training and inference

Thanks for your attention!