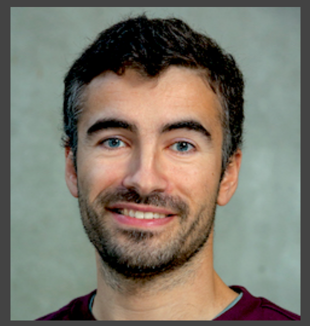
8. Green AI

Sustainable Software Engineering
CS4575


Luís Cruz
L.Cruz@tudelft.nl

Carolin Brandt
C.E.Brandt@tudelft.nl

Enrique Barba Roque
E.BarbaRoque@tudelft.nl

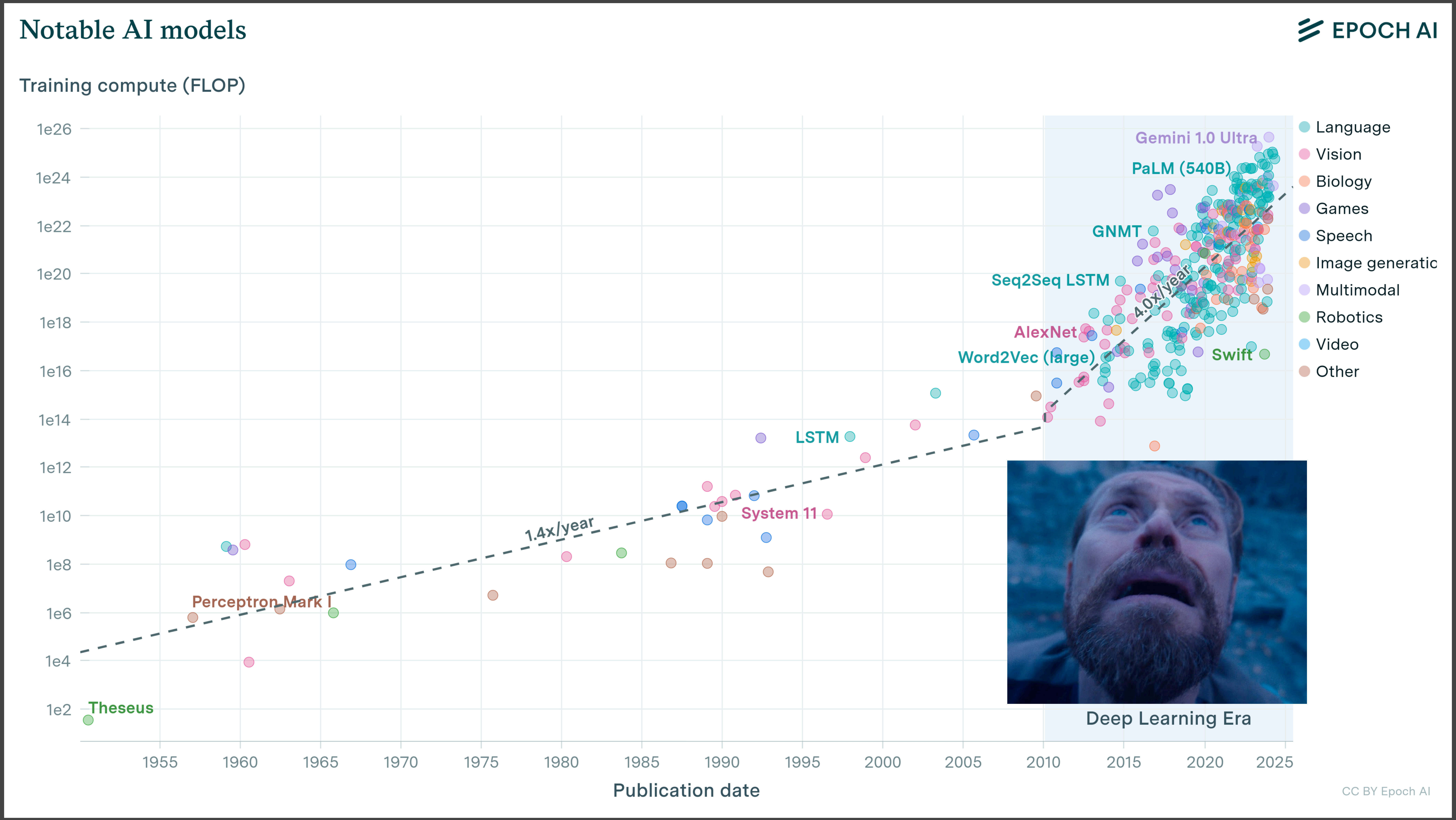
- Overview of Green AI
    - Large language models
- Green data-centric AI
- Model simplification
- Hyper parameter tuning
- Mixture of Experts and DeepSeek
- Green AI at Meta

AI

- Artificial Intelligence (AI) is the branch of computer science that deals with **automating** tasks that typically require **human intelligence**.
- In the past years AI has been widely applied across different domains. E.g., health care, transportation, finance.
- To deploy AI systems, we test them against **benchmarks** (or validation sets).
    - The goal is to outperform the previous existing models.
    - E.g., in Machine Learning we usually resort to accuracy metrics. The highest the accuracy, the better the model.

Since 2012, the amount of computing used for AI training has been doubling every 6 months

- https://epoch.ai/blog/compute-trends



- To create better AI systems we are currently adding
  - **More data**
  - **More experiments**
  - **Larger models**

The Equation of Red AI

$$Cost(R) \propto E \cdot D \cdot H$$

Cost of a single (**E**)xample

Number of (**H**)yperparameters

Size of (**D**)ataset

By Schwartz et al. (2020)

Issues of Red AI

- High costs (hardware, electricity, data access, etc.)
- Limited reproducibility.
- Energy consumption.
- Carbon emissions.
- **SMEs can hardly be competitive.**
- Groundbreaking **AI research is mostly done by tech giants.**

A few examples of Red AI

- Google’s BERT-large
  - 350 million features
  - Trained for 2.5 days using 512 TPU chips, costing $60K+
- Open-GPT3 (now GPT-4/o1)
  - 550 tonnes CO2-eq (Patterson, 2021)
  - 175 billion features
  - API is open but no-pretrained model is available
- AlphaGo
  - 1920 CPUs, 280 GPUs, costing $35M

Red AI in Large Language Models (LLMs)

- **OPT** by Meta reports **75 tons CO2-eq** (1/7 of OpenGPT's footprint). (Also 175billion params)
    - However, **Llama 3** reported **2,290** tons of CO2-eq (7.7M GPU hours training)
    - **Open science**: release includes both the pretrained models and the code needed to train and use them.
- **DeepSeek-V3** claims "only" 2.78M GPU hours
- **Bloom** by Huggingface reports **25 tons**, 51 when considering embodied and operational carbon footprint. (176billion params)

Red AI

Accuracy: 0.999999999

Green AI

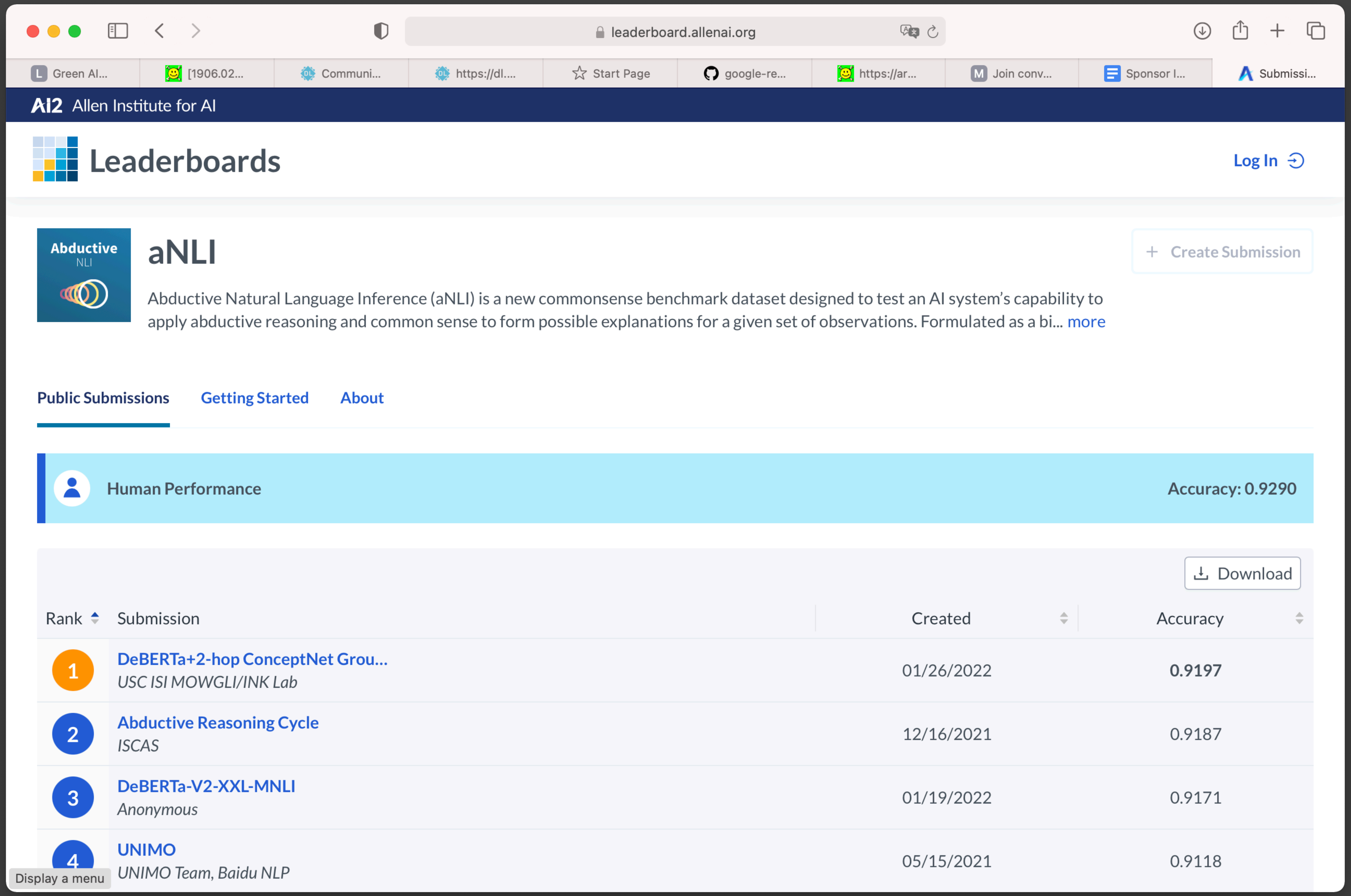
- Energy
- Time
- Reproducibility
- Reusage

How can we adopt Green AI

- **Check whether AI is needed.**
- Select green datacenters.
- Run on **low carbon intensity** hours.
- Opt for **GPU-optimised** solutions (?)
- Opt for **low-power hardware** (e.g., Nvidia Jetson boards)
  - Or GPUs that provide energy metrics (e.g., NVIDIA GPUs via the **nvidia-smi** tool)
- **Report** energy/carbon metrics (e.g., embed in MLFlow?)
- Use pre-trained models (Transfer Learning)
- Preprocess dataset to reduce size.
- Improve parameter-tuning strategy.

Reporting energy/carbon footprint

- We need **benchmarks**.
- AllenAI leaderboard https://leaderboard.allenai.org
  - **No carbon metrics**, yet
- Report comparable proxies for energy consumption.
  - ⚠️ Learning algorithms behave in a non-deterministic
  - ⚠️ Different data-points lead to different energy consumption



aNLI

Abductive Natural Language Inference (aNLI) is a new commonsense benchmark dataset designed to test an AI system's capability to apply abductive reasoning and common sense to form possible explanations for a given set of observations. Formulated as a bi... more

Human Performance Accuracy: 0.9290

Rank	Submission	Created	Accuracy
1	DeBERTa+2-hop ConceptNet Grou... *USC ISI MOWGLI/INK Lab*	01/26/2022	**0.9197**
2	Abductive Reasoning Cycle *ISCAS*	12/16/2021	0.9187
3	DeBERTa-V2-XXL-MNLI *Anonymous*	01/19/2022	0.9171
4	UNIMO *UNIMO Team, Baidu NLP*	05/15/2021	0.9118

Reporting energy/carbon footprint

- Reporting **measured energy consumption**
  - + Accurate
  - + Easy to map to carbon emissions
  - - Hard to measure
  - - Low replicability
- Reporting **time** / estimation based on **time & hardware**
  - + Easy to measure
  - + Correlates with energy consumption in most cases.
  - - Difficult to compare with measurements from other setups
- E.g., **floating point operations** (FPOs) (?)
  - + comparable across different setups
  - + cheap
  - - does not factor in memory energy consumption
  - - does not reflect carbon emissions

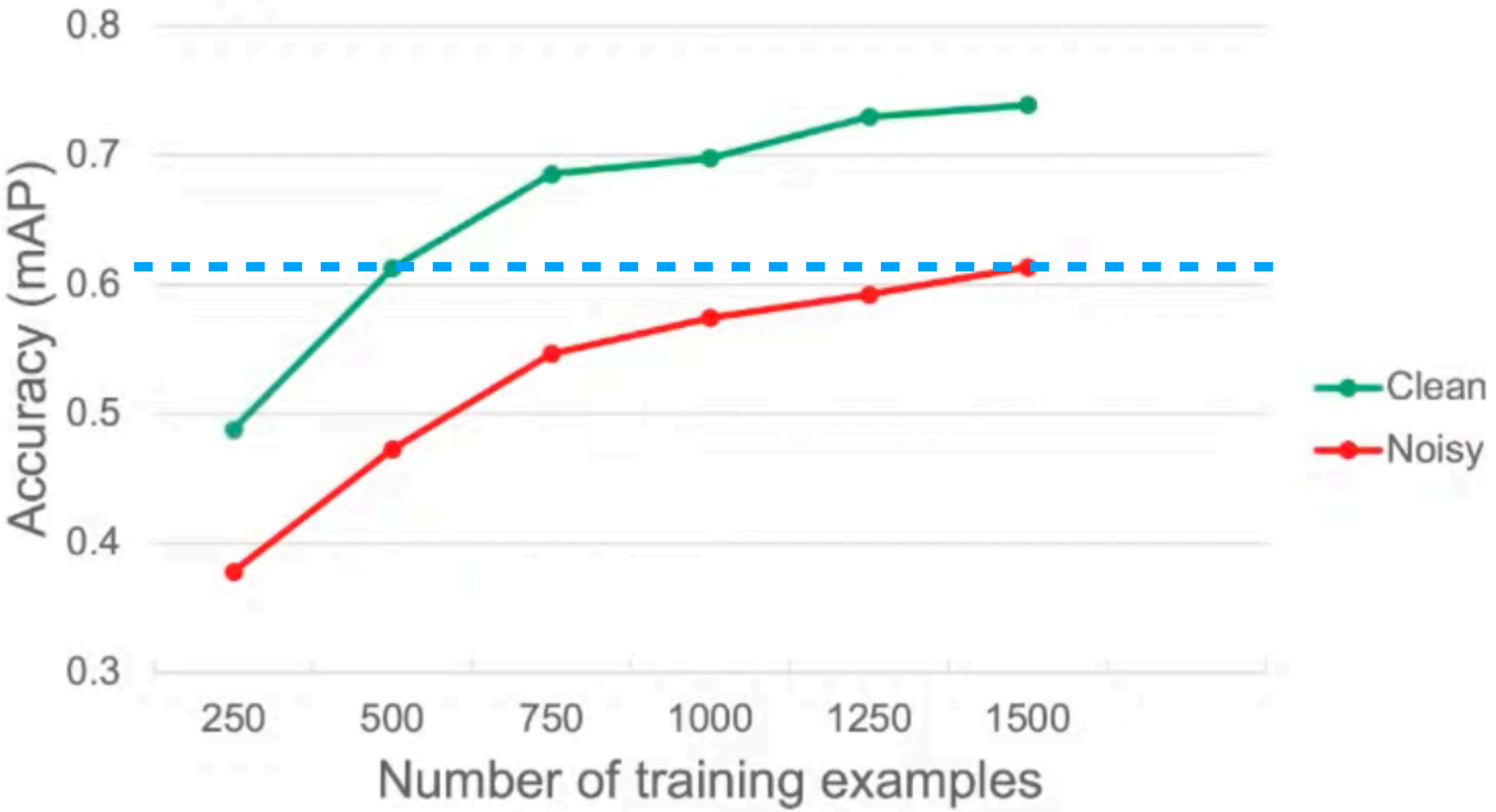
Data-centric AI

Data-centric AI

- Emerging discipline that deals with systematically engineering data to build AI systems.
  - Shift from **improving the training strategy** to **improving the data**.
  - It is better to have **small but reliable** datasets than **large but noisy** datasets.
    - => Improve **data collection**, **data labelling**, and **data preprocessing**.

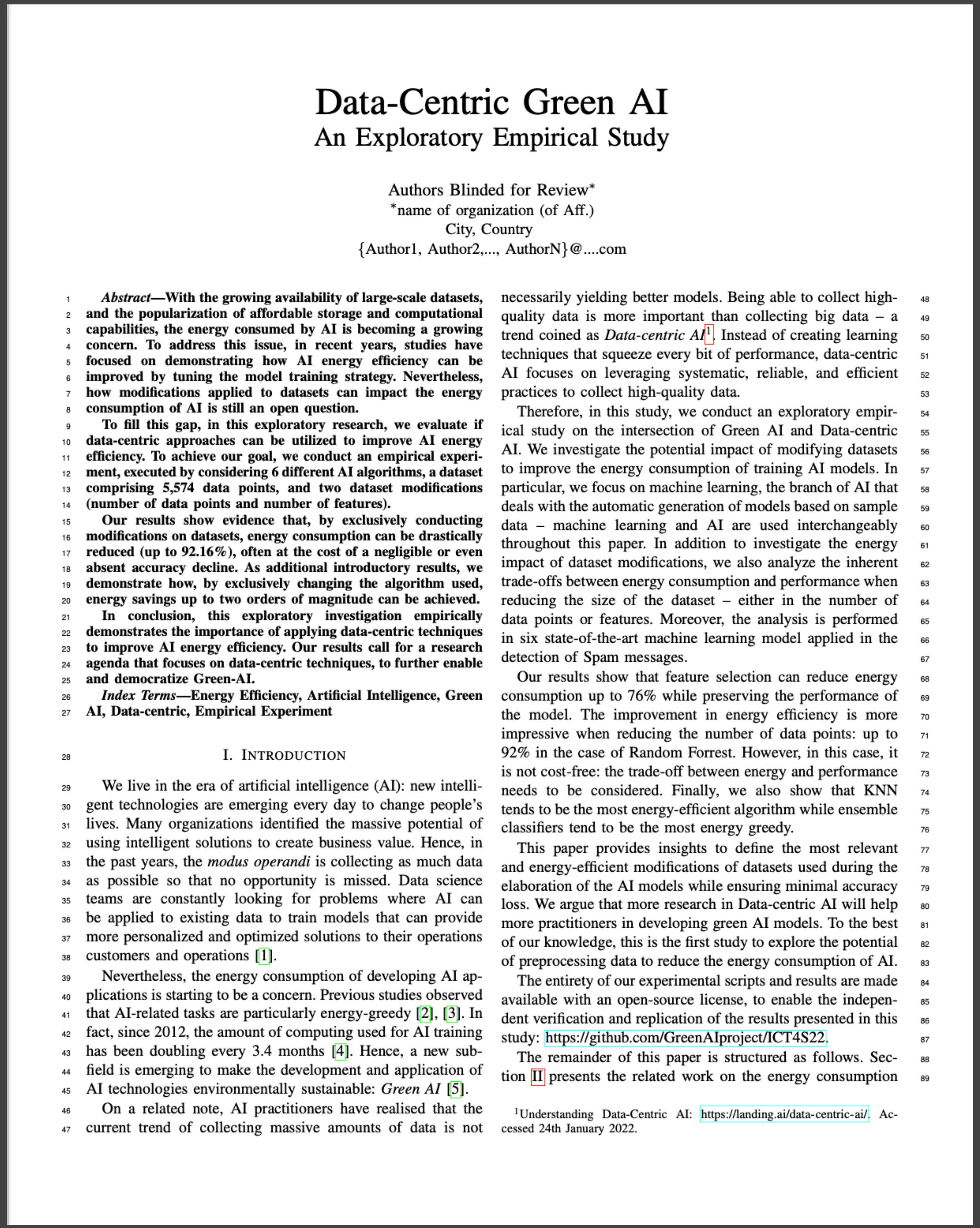
  - More about data-centric AI by Andrew Ng: https://www.youtube.com/watch?v=06-AZXmwHjo

Example: Clean vs. noisy data



Green Data-centric AI

- How do different ML algorithms compare in terms of energy consumption?
- How does **number of rows** relate to the energy consumption of ML models?
- How does **number of features** relate to the energy consumption of ML models?
- What is the impact of reducing data in the **performance** of the model?
- Method -> results -> discussion



Data-Centric Green AI
An Exploratory Empirical Study

Authors Blinded for Review*
*name of organization (of Aff.)
City, Country
{Author1, Author2,..., AuthorN}@....com

***Abstract*—With the growing availability of large-scale datasets, and the popularization of affordable storage and computational capabilities, the energy consumed by AI is becoming a growing concern. To address this issue, in recent years, studies have focused on demonstrating how AI energy efficiency can be improved by tuning the model training strategy. Nevertheless, how modifications applied to datasets can impact the energy consumption of AI is still an open question.**

To fill this gap, in this exploratory research, we evaluate if data-centric approaches can be utilized to improve AI energy efficiency. To achieve our goal, we conduct an empirical experiment, executed by considering 6 different AI algorithms, a dataset comprising 5,574 data points, and two dataset modifications (number of data points and number of features).

Our results show evidence that, by exclusively conducting modifications on datasets, energy consumption can be drastically reduced (up to 92.16%), often at the cost of a negligible or even absent accuracy decline. As additional introductory results, we demonstrate how, by exclusively changing the algorithm used, energy savings up to two orders of magnitude can be achieved.

In conclusion, this exploratory investigation empirically demonstrates the importance of applying data-centric techniques to improve AI energy efficiency. Our results call for a research agenda that focuses on data-centric techniques, to further enable and democratize Green-AI.

***Index Terms*—Energy Efficiency, Artificial Intelligence, Green AI, Data-centric, Empirical Experiment**

I. Introduction

We live in the era of artificial intelligence (AI): new intelligent technologies are emerging every day to change people's lives. Many organizations identified the massive potential of using intelligent solutions to create business value. Hence, in the past years, the *modus operandi* is collecting as much data as possible so that no opportunity is missed. Data science teams are constantly looking for problems where AI can be applied to existing data to train models that can provide more personalized and optimized solutions to their operations customers and operations [1].

Nevertheless, the energy consumption of developing AI applications is starting to be a concern. Previous studies observed that AI-related tasks are particularly energy-greedy [2], [3]. In fact, since 2012, the amount of computing used for AI training has been doubling every 3.4 months [4]. Hence, a new subfield is emerging to make the development and application of AI technologies environmentally sustainable: *Green AI* [5].

On a related note, AI practitioners have realised that the current trend of collecting massive amounts of data is not necessarily yielding better models. Being able to collect high-quality data is more important than collecting big data – a trend coined as *Data-centric AI*[1]. Instead of creating learning techniques that squeeze every bit of performance, data-centric AI focuses on leveraging systematic, reliable, and efficient practices to collect high-quality data.

Therefore, in this study, we conduct an exploratory empirical study on the intersection of Green AI and Data-centric AI. We investigate the potential impact of modifying datasets to improve the energy consumption of training AI models. In particular, we focus on machine learning, the branch of AI that deals with the automatic generation of models based on sample data – machine learning and AI are used interchangeably throughout this paper. In addition to investigate the energy impact of dataset modifications, we also analyze the inherent trade-offs between energy consumption and performance when reducing the size of the dataset – either in the number of data points or features. Moreover, the analysis is performed in six state-of-the-art machine learning model applied in the detection of Spam messages.

Our results show that feature selection can reduce energy consumption up to 76% while preserving the performance of the model. The improvement in energy efficiency is more impressive when reducing the number of data points: up to 92% in the case of Random Forrest. However, in this case, it is not cost-free: the trade-off between energy and performance needs to be considered. Finally, we also show that KNN tends to be the most energy-efficient algorithm while ensemble classifiers tend to be the most energy greedy.

This paper provides insights to define the most relevant and energy-efficient modifications of datasets used during the elaboration of the AI models while ensuring minimal accuracy loss. We argue that more research in Data-centric AI will help more practitioners in developing green AI models. To the best of our knowledge, this is the first study to explore the potential of preprocessing data to reduce the energy consumption of AI.

The entirety of our experimental scripts and results are made available with an open-source license, to enable the independent verification and replication of the results presented in this study: https://github.com/GreenAIproject/ICT4S22.

The remainder of this paper is structured as follows. Section II presents the related work on the energy consumption

[1] Understanding Data-Centric AI: https://landing.ai/data-centric-ai/. Accessed 24th January 2022.

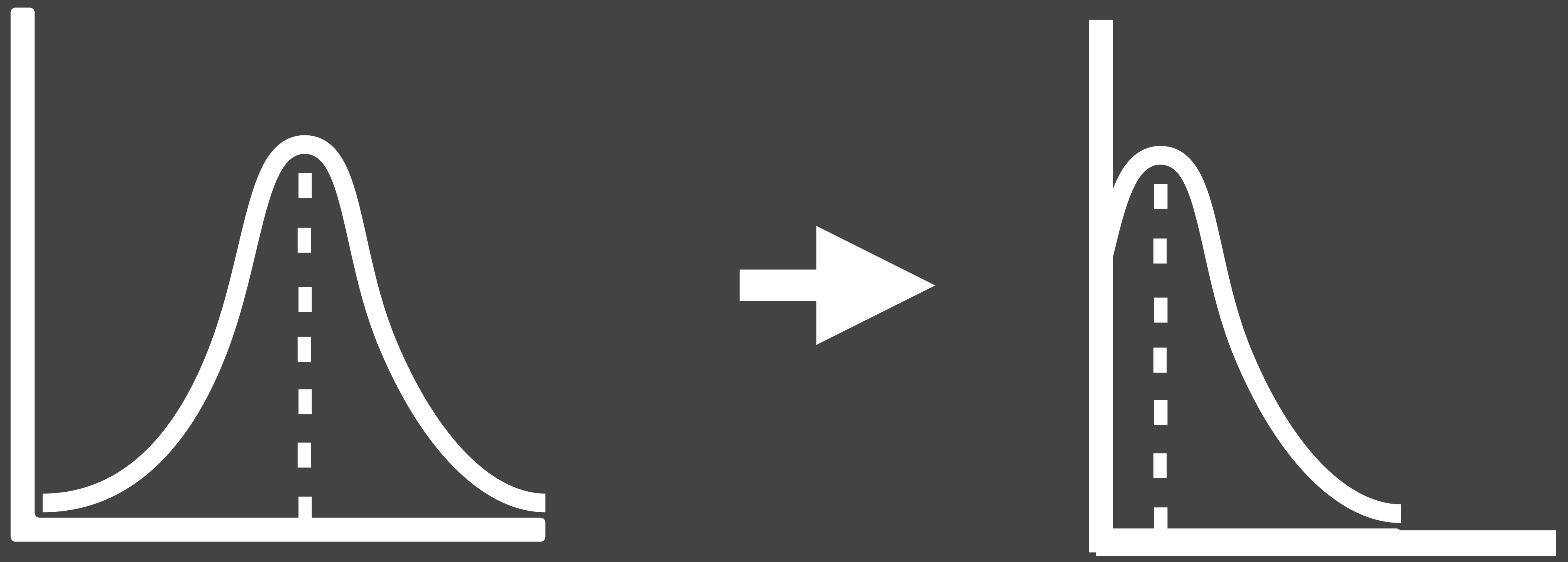
Method

- Single object of study: natural language model to **detect spam messages**.
- 6 machine learning algorithms: **SVM, Decision Tree, KNN, Random Forrest, AdaBoost, Bagging Classifier**.
- Reduce the number of rows. 10%, 20%, .., 100%
    - **Stratified random sampling** (?)
- Reduce the number of features. 10%, 20%, .., 100%
    - **Feature importance** metric based on the Chi-Square Test (Chi2)
- Estimate energy consumption using a RAPL-based tool. (?)

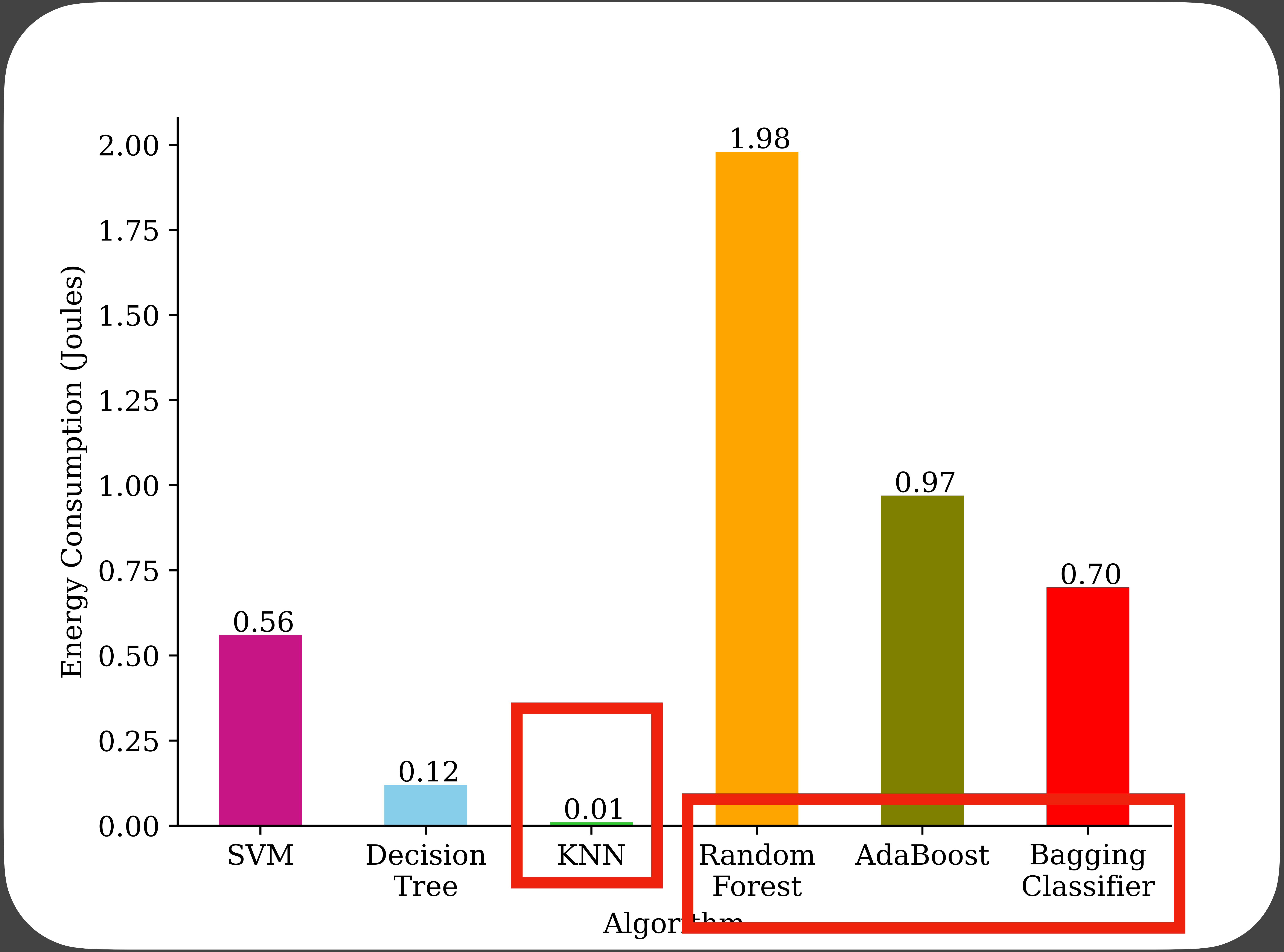
- Repeat 30 times
- Fix random seeds
- …

(?)

- Data was **not Normal** => **tailed Normal distribution**.



Results: energy consumption of algorithms



Results: energy vs data shape

SVM
Decision Tree
KNN
Random Forest
AdaBoost
Bagging Classifier
Energy Consumption (Joules)
Number of Datapoints
Energy Consumption (Joules)
Number of Features
0.0
0.1
0.2
0.3
0.4
0.5
1000
2000
3000
4000
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.000
0.002
0.004
0.006
0.008
0.010
0.5
1.0
1.5
2.0
0.2
0.4
0.6
0.8
0.7
2000
4000
6000
8000

Results: performance vs data shape

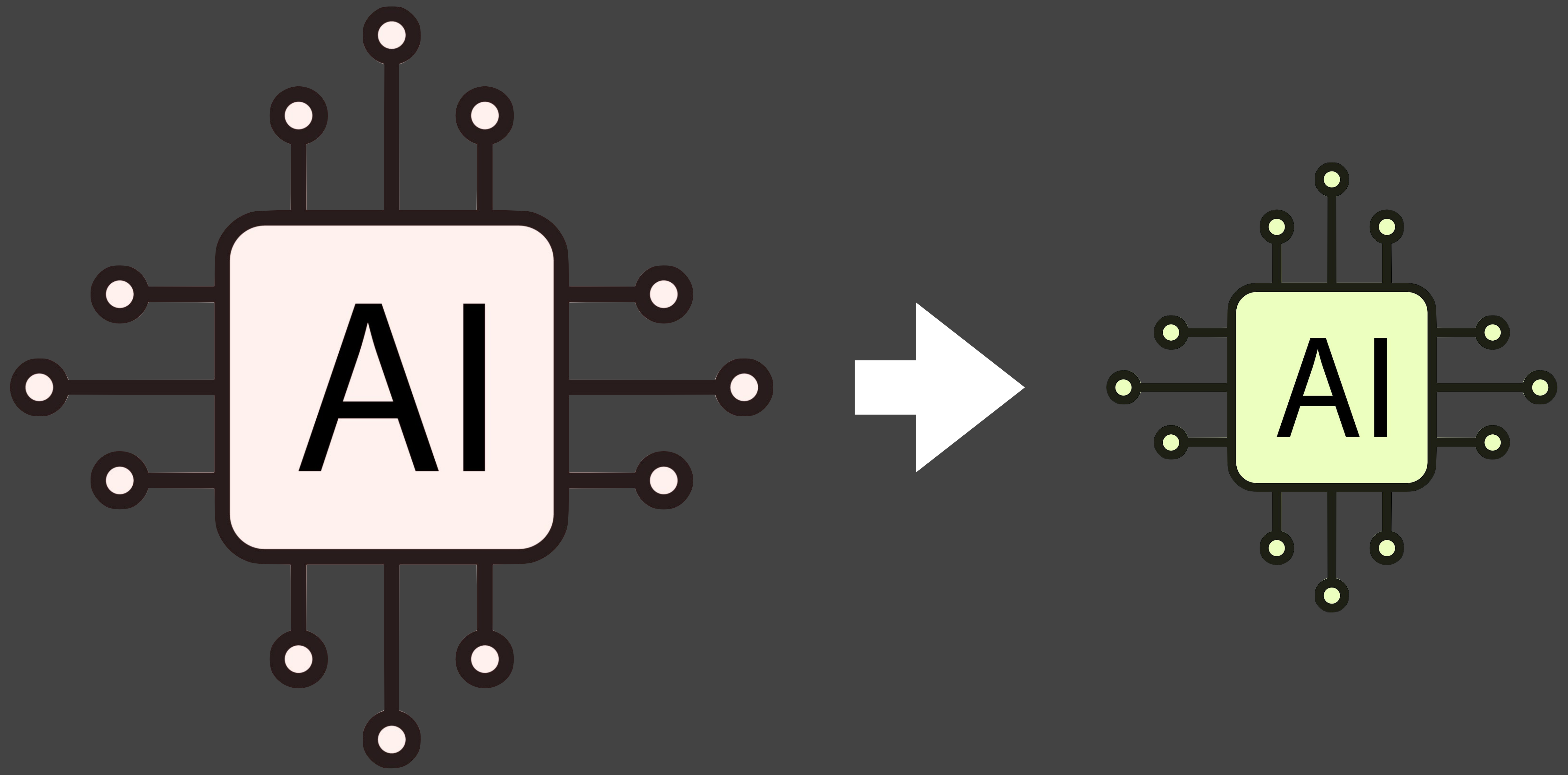
SVM
Decision Tree
KNN
Random Forest
AdaBoost
Bagging Classifier
F1-score
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
Number of Datapoints
SVM
Decision Tree
KNN
Random Forest
AdaBoost
Bagging Classifier
F1-score
0
2000
4000
6000
8000
Number of Features

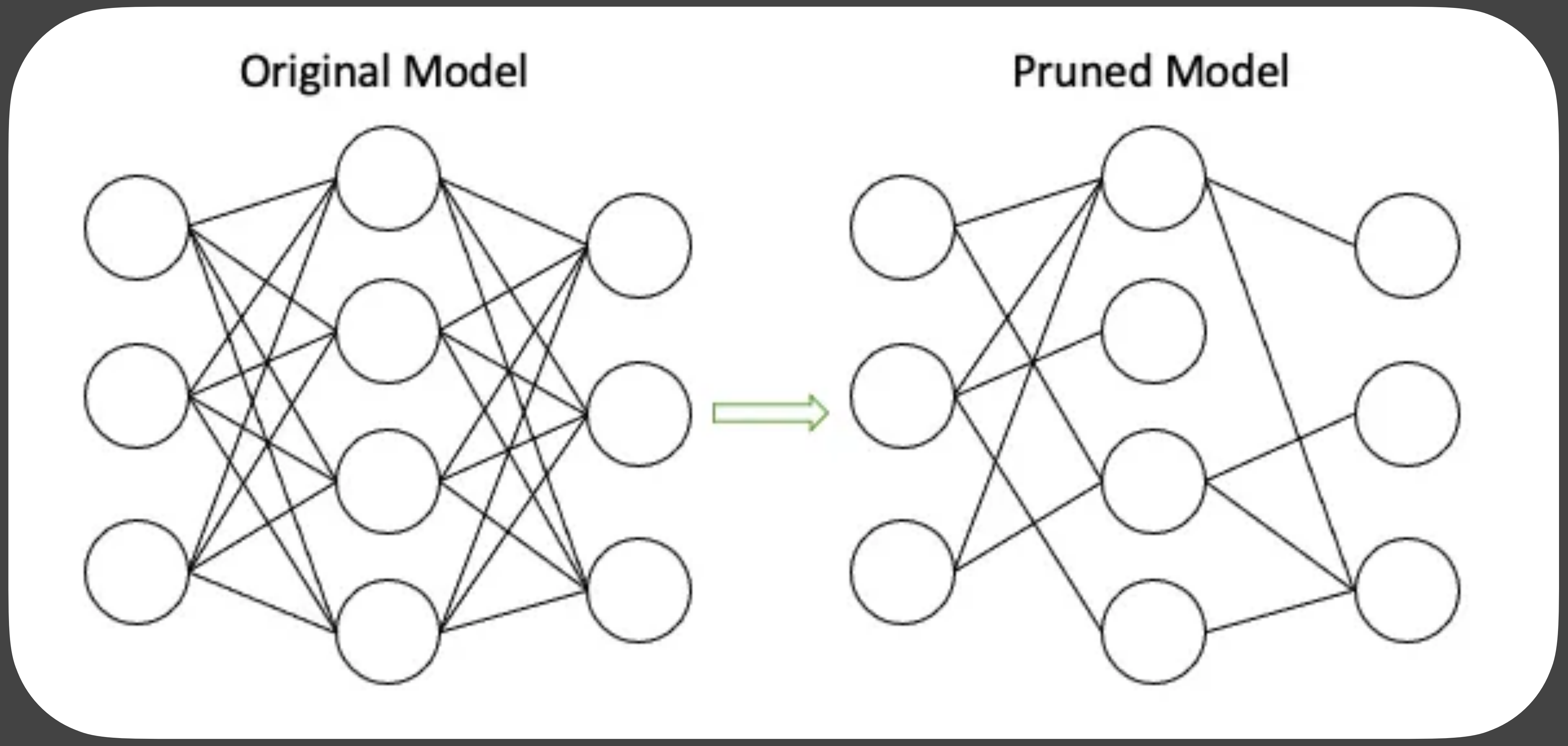
Discussion

- Other data properties should be investigated.
  - E.g., data types
- **Reporting energy data** is essential. It can lead to different model selection without hindering model performance.
- There is a big opportunity in **Model and Data Simplification.**

Data/Model Simplification

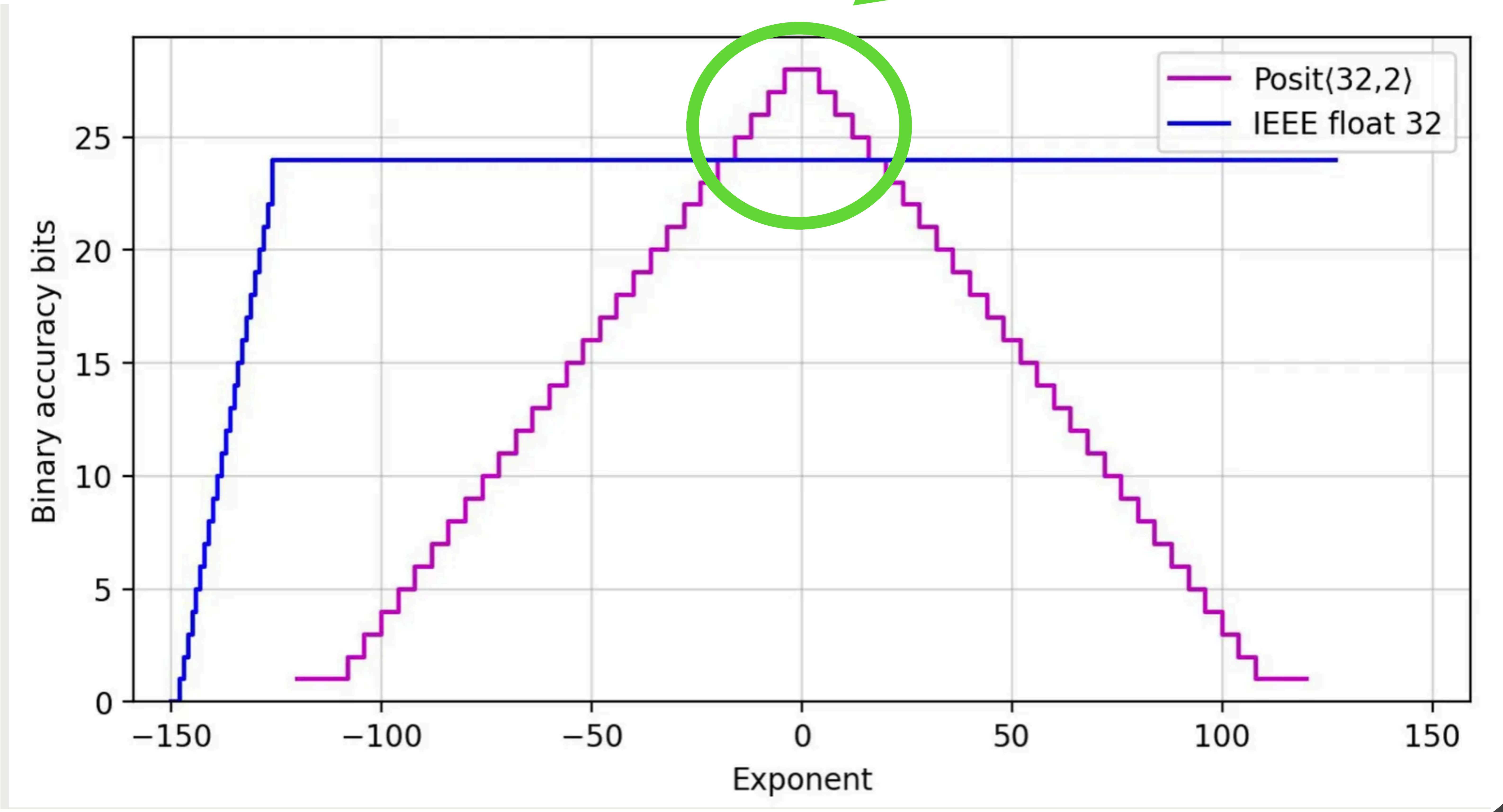
- (?)
- Data selection
- Data quantisation. **Posit?**
- Data distillation
- Coreset extraction (?)
- Model distillation
- Model quantisation
- Model pruning
- …





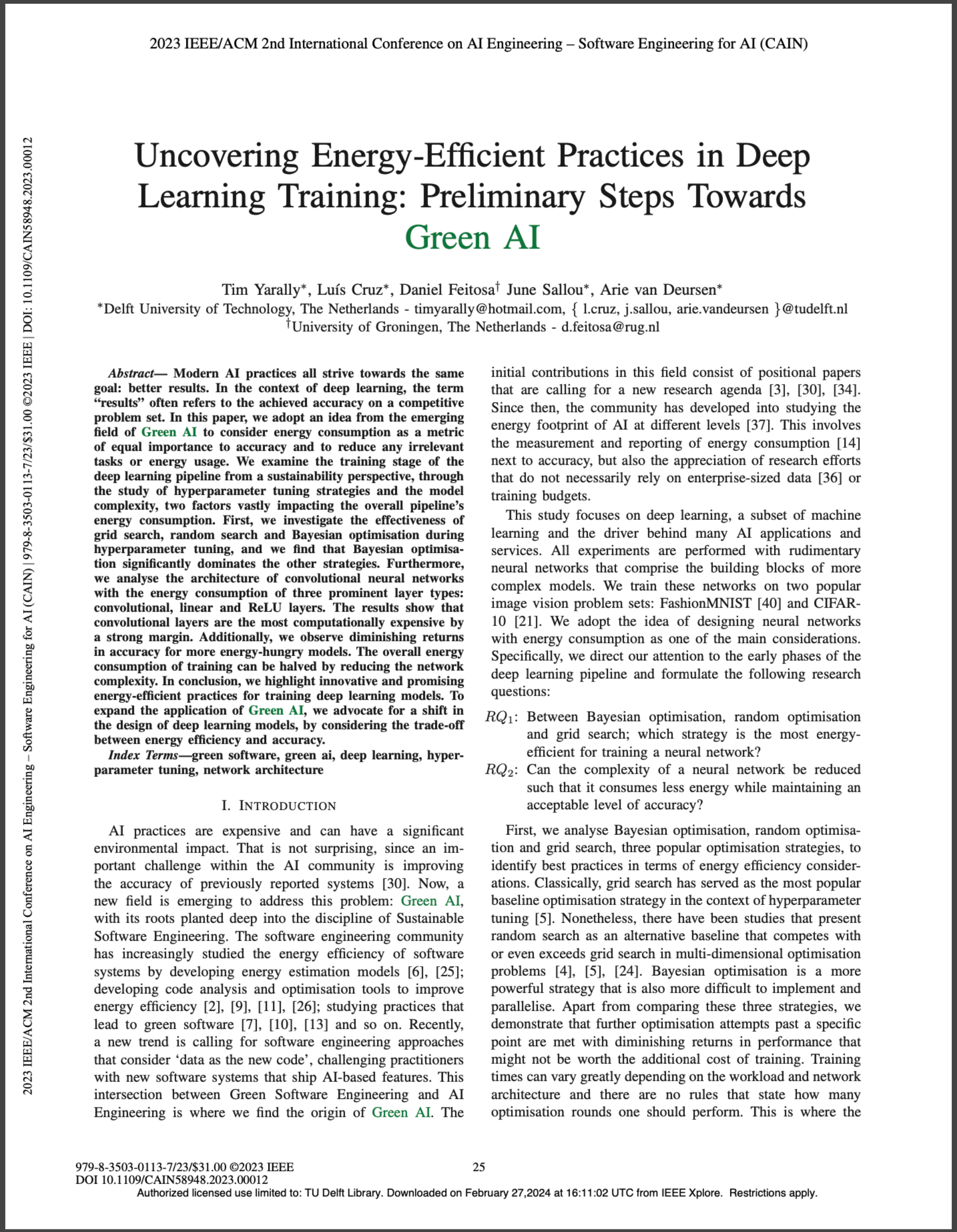
Posit vs Float

Better for DL use cases



More about this: https://spectrum.ieee.org/floating-point-numbers-posits-processor

How can we tune learning parameters efficiently?



2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)

2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN) | 979-8-3503-0113-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAIN58948.2023.00012

Uncovering Energy-Efficient Practices in Deep Learning Training: Preliminary Steps Towards Green AI

Tim Yarally*, Luís Cruz*, Daniel Feitosa† June Sallou*, Arie van Deursen*
*Delft University of Technology, The Netherlands - timyarally@hotmail.com, { l.cruz, j.sallou, arie.vandeursen }@tudelft.nl
†University of Groningen, The Netherlands - d.feitosa@rug.nl

***Abstract*— Modern AI practices all strive towards the same goal: better results. In the context of deep learning, the term "results" often refers to the achieved accuracy on a competitive problem set. In this paper, we adopt an idea from the emerging field of Green AI to consider energy consumption as a metric of equal importance to accuracy and to reduce any irrelevant tasks or energy usage. We examine the training stage of the deep learning pipeline from a sustainability perspective, through the study of hyperparameter tuning strategies and the model complexity, two factors vastly impacting the overall pipeline's energy consumption. First, we investigate the effectiveness of grid search, random search and Bayesian optimisation during hyperparameter tuning, and we find that Bayesian optimisation significantly dominates the other strategies. Furthermore, we analyse the architecture of convolutional neural networks with the energy consumption of three prominent layer types: convolutional, linear and ReLU layers. The results show that convolutional layers are the most computationally expensive by a strong margin. Additionally, we observe diminishing returns in accuracy for more energy-hungry models. The overall energy consumption of training can be halved by reducing the network complexity. In conclusion, we highlight innovative and promising energy-efficient practices for training deep learning models. To expand the application of Green AI, we advocate for a shift in the design of deep learning models, by considering the trade-off between energy efficiency and accuracy.**

***Index Terms*—green software, green ai, deep learning, hyperparameter tuning, network architecture**

I. Introduction

AI practices are expensive and can have a significant environmental impact. That is not surprising, since an important challenge within the AI community is improving the accuracy of previously reported systems [30]. Now, a new field is emerging to address this problem: Green AI, with its roots planted deep into the discipline of Sustainable Software Engineering. The software engineering community has increasingly studied the energy efficiency of software systems by developing energy estimation models [6], [25]; developing code analysis and optimisation tools to improve energy efficiency [2], [9], [11], [26]; studying practices that lead to green software [7], [10], [13] and so on. Recently, a new trend is calling for software engineering approaches that consider 'data as the new code', challenging practitioners with new software systems that ship AI-based features. This intersection between Green Software Engineering and AI Engineering is where we find the origin of Green AI. The initial contributions in this field consist of positional papers that are calling for a new research agenda [3], [30], [34]. Since then, the community has developed into studying the energy footprint of AI at different levels [37]. This involves the measurement and reporting of energy consumption [14] next to accuracy, but also the appreciation of research efforts that do not necessarily rely on enterprise-sized data [36] or training budgets.

This study focuses on deep learning, a subset of machine learning and the driver behind many AI applications and services. All experiments are performed with rudimentary neural networks that comprise the building blocks of more complex models. We train these networks on two popular image vision problem sets: FashionMNIST [40] and CIFAR-10 [21]. We adopt the idea of designing neural networks with energy consumption as one of the main considerations. Specifically, we direct our attention to the early phases of the deep learning pipeline and formulate the following research questions:

RQ_1: Between Bayesian optimisation, random optimisation and grid search; which strategy is the most energy-efficient for training a neural network?

RQ_2: Can the complexity of a neural network be reduced such that it consumes less energy while maintaining an acceptable level of accuracy?

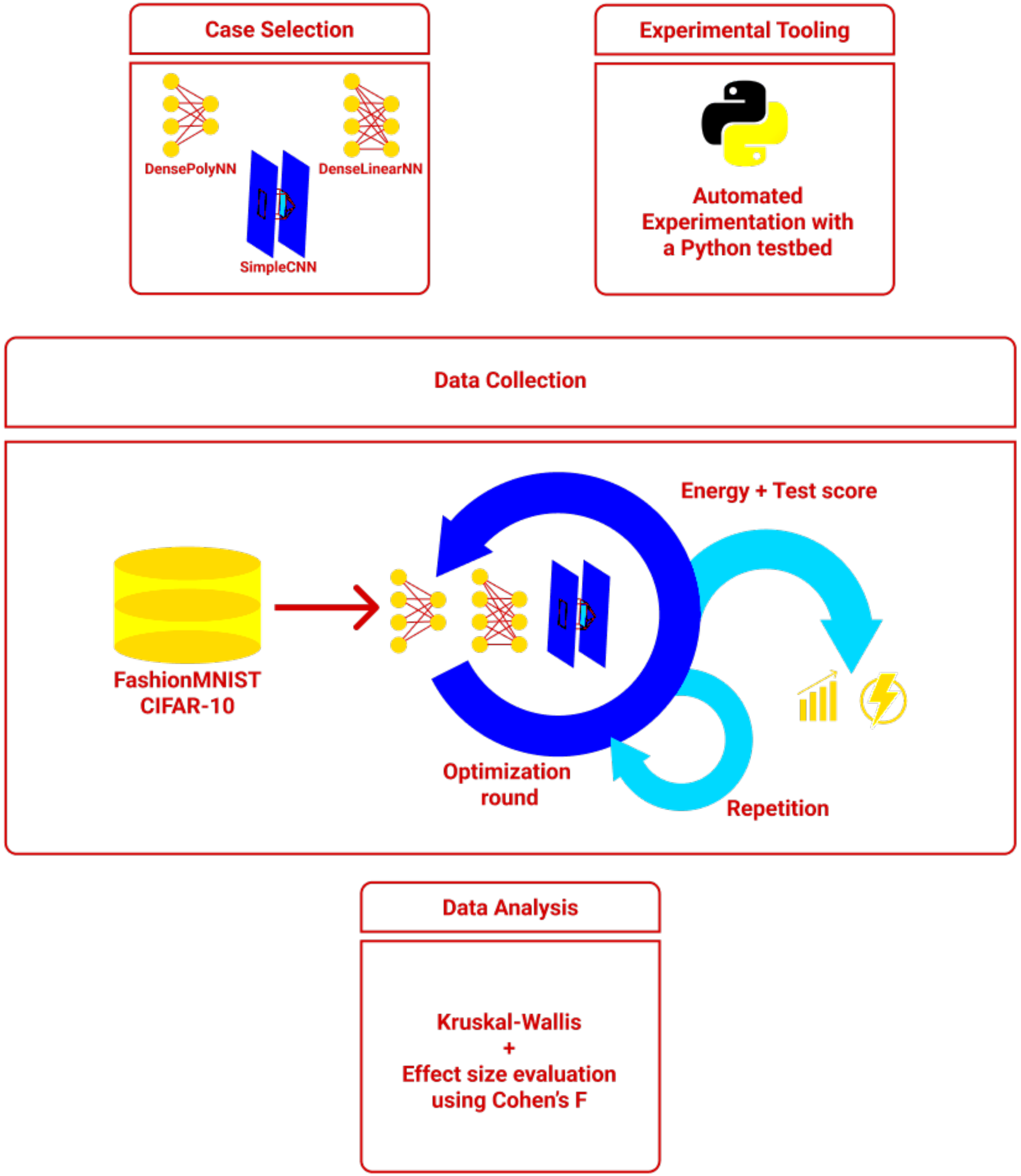
First, we analyse Bayesian optimisation, random optimisation and grid search, three popular optimisation strategies, to identify best practices in terms of energy efficiency considerations. Classically, grid search has served as the most popular baseline optimisation strategy in the context of hyperparameter tuning [5]. Nonetheless, there have been studies that present random search as an alternative baseline that competes with or even exceeds grid search in multi-dimensional optimisation problems [4], [5], [24]. Bayesian optimisation is a more powerful strategy that is also more difficult to implement and parallelise. Apart from comparing these three strategies, we demonstrate that further optimisation attempts past a specific point are met with diminishing returns in performance that might not be worth the additional cost of training. Training times can vary greatly depending on the workload and network architecture and there are no rules that state how many optimisation rounds one should perform. This is where the

979-8-3503-0113-7/23/$31.00 ©2023 IEEE
DOI 10.1109/CAIN58948.2023.00012

Tim's MSc Thesis, 2022
Accepted at CAIN'23

Hyper parameter tuning

- When training an ML model, there are several **parameters** that need to be **tuned**.
  - E.g., in SVM we have the *Regularization parameter* C, the kernel function, the degree of the kernel function, and depending on the case, many other.
  - The common approach revolves around **grid search**. The user provides a sequence of possible values for each parameter and the pipeline runs **all possible combinations**.
    - **Our question:** Can we save energy with alternative approaches?
    - We studied **Grid Search**, **Random Search** and **Bayesian Optimisation**.



Case Selection

Experimental Tooling

Automated Experimentation with a Python testbed

Data Collection

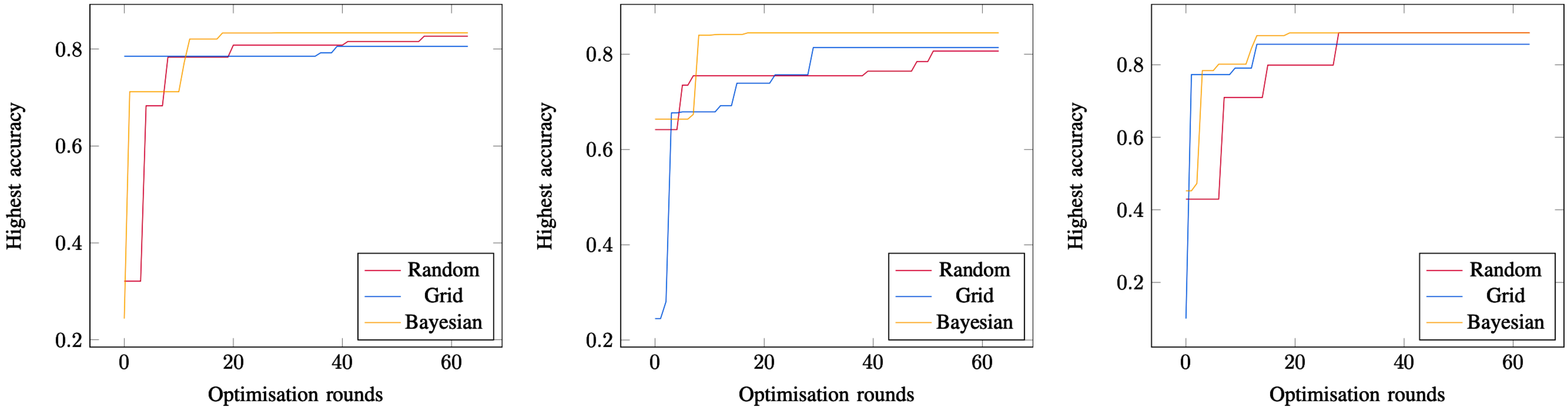
Data Analysis

Kruskal-Wallis
+
Effect size evaluation using Cohen's F

Results

Conclusions?

- **Bayesian** converges faster.
- No clear winner between Grid and Random



(a) DensePolyNN

(b) DenseLinearNN

(c) SimpleCNN

DeepSeekMoE

Preprint at ArXiv, 2024 https://arxiv.org/pdf/2401.06066

deepseek

DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models

Damai Dai*[1,2], Chengqi Deng[1], Chenggang Zhao*[1,3], R.X. Xu[1], Huazuo Gao[1], Deli Chen[1], Jiashi Li[1], Wangding Zeng[1], Xingkai Yu*[1,4], Y. Wu[1], Zhenda Xie[1], Y.K. Li[1], Panpan Huang[1], Fuli Luo[1], Chong Ruan[1], Zhifang Sui[2], Wenfeng Liang[1]

[1]**DeepSeek-AI**
[2]**National Key Laboratory for Multimedia Information Processing, Peking University**
[3]**Institute for Interdisciplinary Information Sciences, Tsinghua University**
[4]**National Key Laboratory for Novel Software Technology, Nanjing University**
{daidamai, szf}@pku.edu.cn, {wenfeng.liang}@deepseek.com
https://github.com/deepseek-ai/DeepSeek-MoE

Abstract

In the era of large language models, Mixture-of-Experts (MoE) is a promising architecture for managing computational costs when scaling up model parameters. However, conventional MoE architectures like GShard, which activate the top-K out of N experts, face challenges in ensuring expert specialization, i.e. each expert acquires non-overlapping and focused knowledge. In response, we propose the **DeepSeekMoE** architecture towards ultimate expert specialization. It involves two principal strategies: (1) finely segmenting the experts into mN ones and activating mK from them, allowing for a more flexible combination of activated experts; (2) isolating K_s experts as shared ones, aiming at capturing common knowledge and mitigating redundancy in routed experts. Starting from a modest scale with 2B parameters, we demonstrate that DeepSeekMoE 2B achieves comparable performance with GShard 2.9B, which has 1.5× expert parameters and computation. In addition, DeepSeekMoE 2B nearly approaches the performance of its dense counterpart with the same number of total parameters, which set the upper bound of MoE models. Subsequently, we scale up DeepSeekMoE to 16B parameters and show that it achieves comparable performance with LLaMA2 7B, with only about 40% of computations. Further, our preliminary efforts to scale up DeepSeekMoE to 145B parameters consistently validate its substantial advantages over the GShard architecture, and show its performance comparable with DeepSeek 67B, using only 28.5% (maybe even 18.2%) of computations.

1. Introduction

Recent research and practices have empirically demonstrated that, with sufficient training data available, scaling language models with increased parameters and computational budgets can yield remarkably stronger models (Brown et al., 2020; Hoffmann et al., 2022; OpenAI, 2023; Touvron et al., 2023a). It is imperative to acknowledge, however, that the endeavor to scale models to an extremely large scale is also associated with exceedingly high computational costs. Considering the substantial costs, the Mixture-of-Experts (MoE) architecture (Jacobs et al., 1991; Jordan and Jacobs, 1994; Shazeer et al., 2017) has emerged as a popular solution. It can

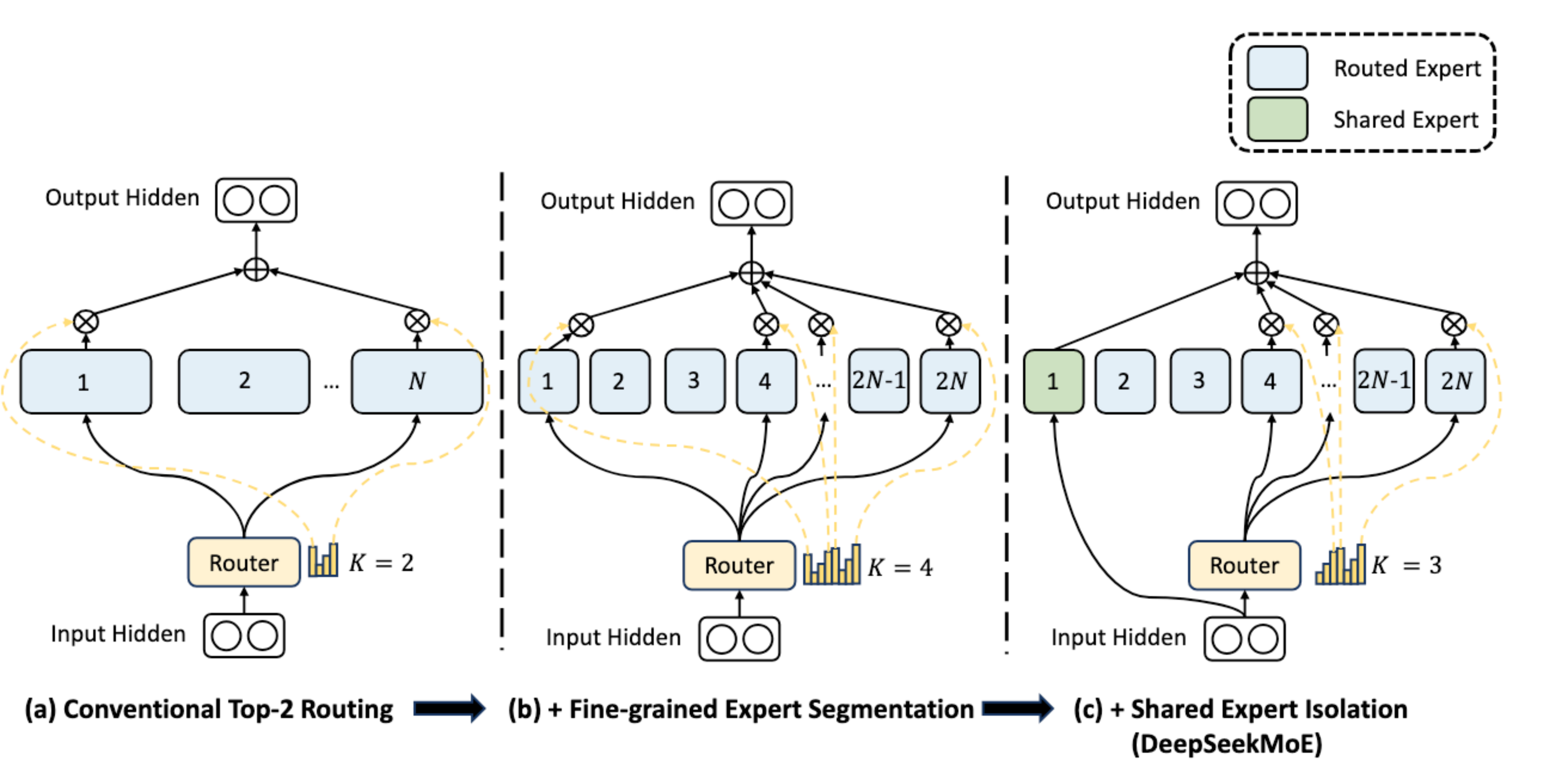
*Contribution during internship at DeepSeek-AI.

Mixture of Experts

- Llama3.1 has **405B** parameters, DeepSeek V3 **671B**
  - Yet DeepSeek has quicker inference times and claims less energy consumption (?)

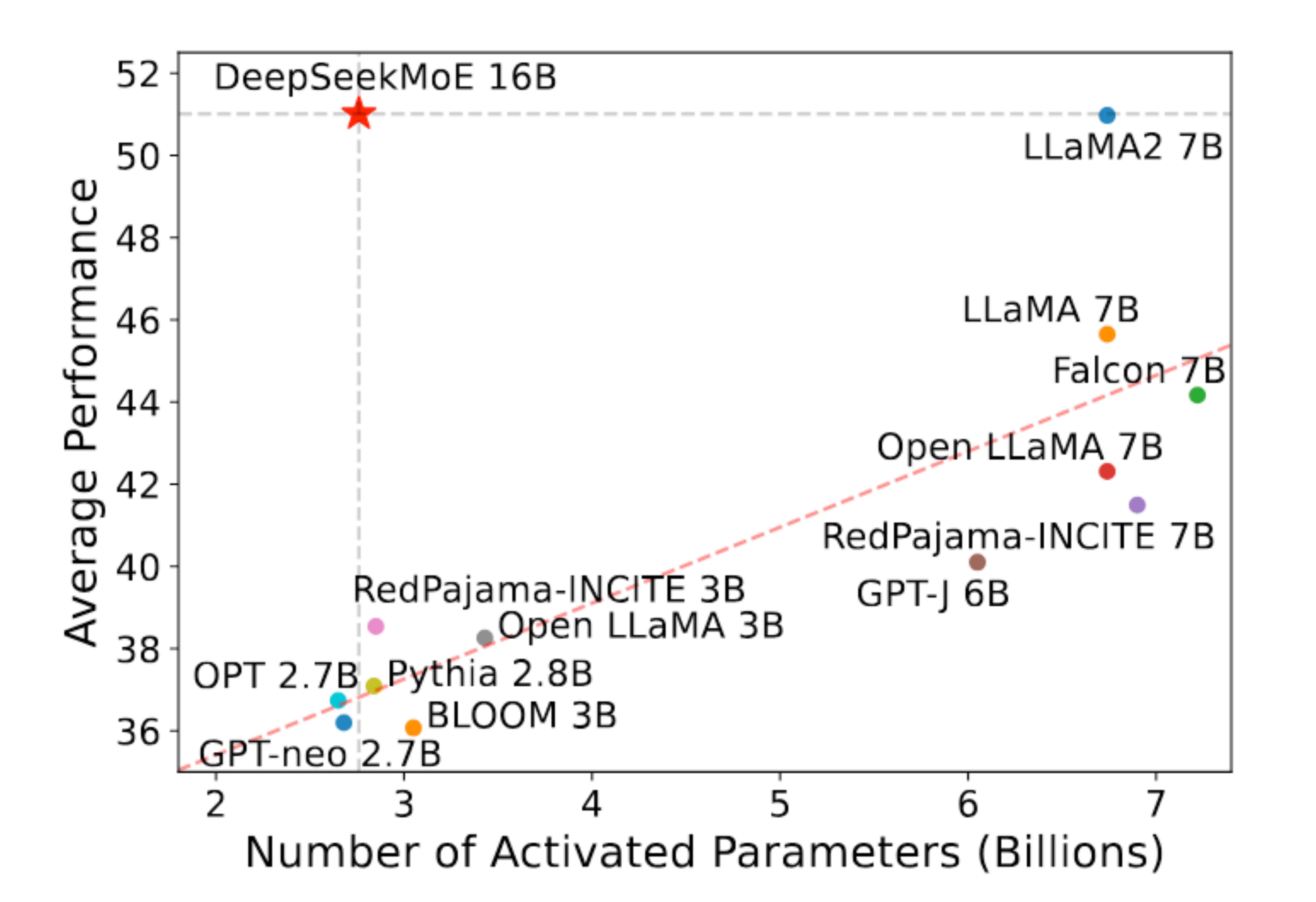
- Divide the model into smaller blocks of **experts**
- Tokens get routed to certain experts based on the query
- Only part of the network is active during inference
  - DeepSeek claims only **37B out of 671B parameters** get active

DeepSeekMoE



(a) Conventional Top-2 Routing → (b) + Fine-grained Expert Segmentation → (c) + Shared Expert Isolation (DeepSeekMoE)

DeepSeekMoE



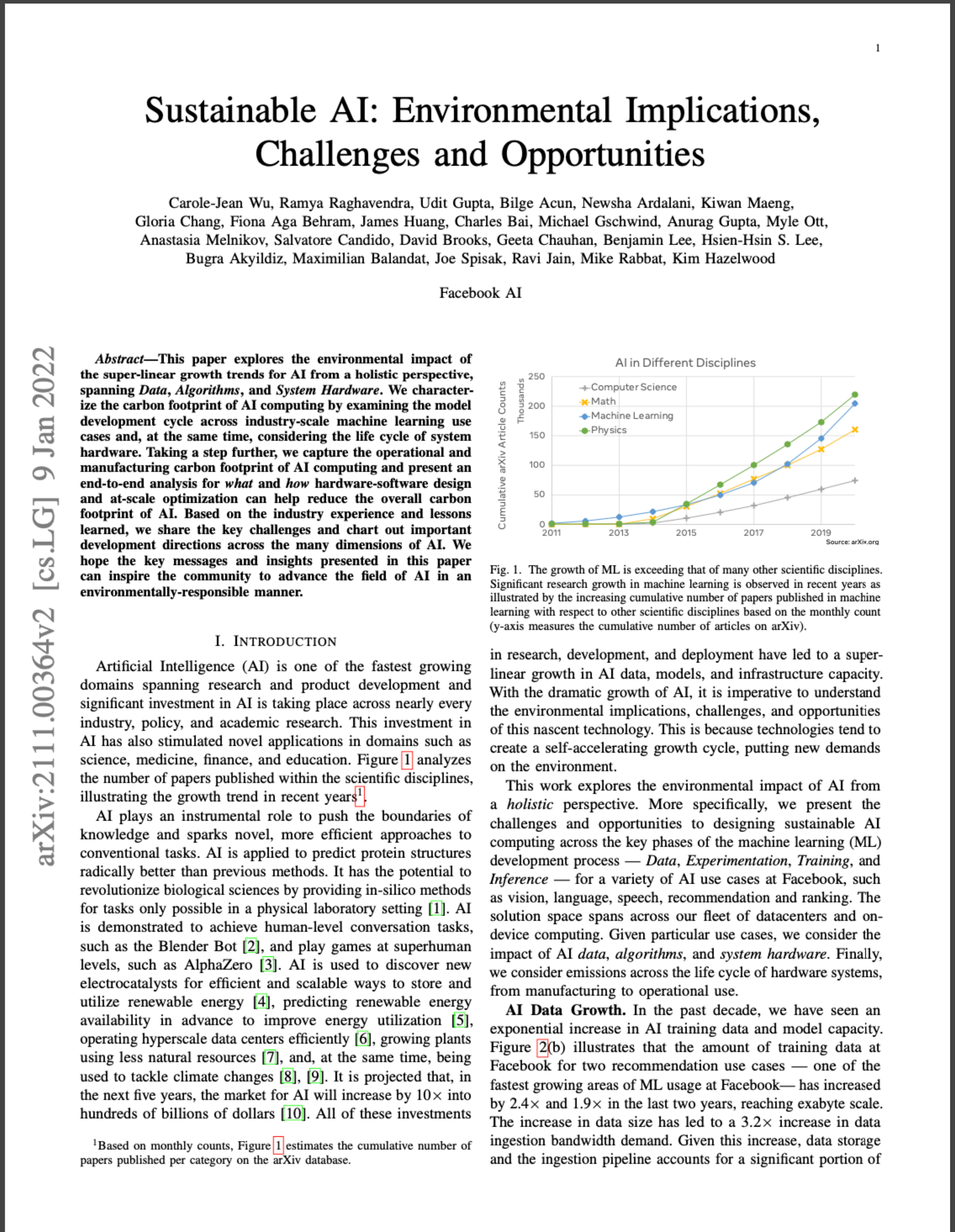
- Comparable performance to LLaMA2 7B effectively using less half the parameters

- Less computational power

- **Problems** (?)
  - Still need to load all the parameters
  - High memory -> **high embodied carbon**

Green AI at ~~Facebook~~Meta

Sustainable AI: Environmental Implications, Challenges and Opportunities (2022)



Sustainable AI: Environmental Implications, Challenges and Opportunities

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng,
Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott,
Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee,
Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, Kim Hazelwood

Facebook AI

***Abstract*—This paper explores the environmental impact of the super-linear growth trends for AI from a holistic perspective, spanning *Data*, *Algorithms*, and *System Hardware*. We characterize the carbon footprint of AI computing by examining the model development cycle across industry-scale machine learning use cases and, at the same time, considering the life cycle of system hardware. Taking a step further, we capture the operational and manufacturing carbon footprint of AI computing and present an end-to-end analysis for *what* and *how* hardware-software design and at-scale optimization can help reduce the overall carbon footprint of AI. Based on the industry experience and lessons learned, we share the key challenges and chart out important development directions across the many dimensions of AI. We hope the key messages and insights presented in this paper can inspire the community to advance the field of AI in an environmentally-responsible manner.**

I. Introduction

Artificial Intelligence (AI) is one of the fastest growing domains spanning research and product development and significant investment in AI is taking place across nearly every industry, policy, and academic research. This investment in AI has also stimulated novel applications in domains such as science, medicine, finance, and education. Figure 1 analyzes the number of papers published within the scientific disciplines, illustrating the growth trend in recent years[1].

AI plays an instrumental role to push the boundaries of knowledge and sparks novel, more efficient approaches to conventional tasks. AI is applied to predict protein structures radically better than previous methods. It has the potential to revolutionize biological sciences by providing in-silico methods for tasks only possible in a physical laboratory setting [1]. AI is demonstrated to achieve human-level conversation tasks, such as the Blender Bot [2], and play games at superhuman levels, such as AlphaZero [3]. AI is used to discover new electrocatalysts for efficient and scalable ways to store and utilize renewable energy [4], predicting renewable energy availability in advance to improve energy utilization [5], operating hyperscale data centers efficiently [6], growing plants using less natural resources [7], and, at the same time, being used to tackle climate changes [8], [9]. It is projected that, in the next five years, the market for AI will increase by 10× into hundreds of billions of dollars [10]. All of these investments in research, development, and deployment have led to a super-linear growth in AI data, models, and infrastructure capacity. With the dramatic growth of AI, it is imperative to understand the environmental implications, challenges, and opportunities of this nascent technology. This is because technologies tend to create a self-accelerating growth cycle, putting new demands on the environment.

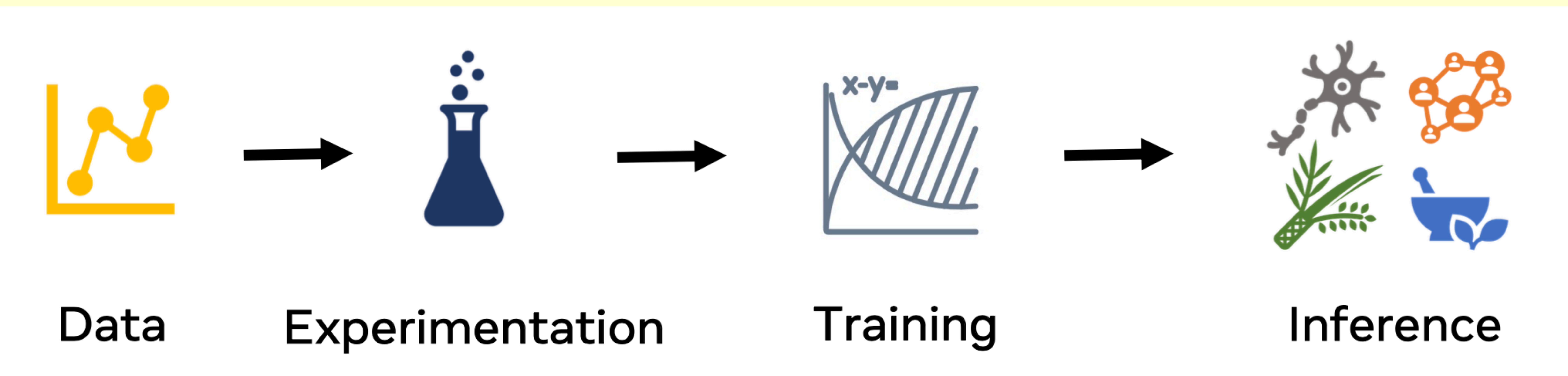
Fig. 1. The growth of ML is exceeding that of many other scientific disciplines. Significant research growth in machine learning is observed in recent years as illustrated by the increasing cumulative number of papers published in machine learning with respect to other scientific disciplines based on the monthly count (y-axis measures the cumulative number of articles on arXiv).

This work explores the environmental impact of AI from a *holistic* perspective. More specifically, we present the challenges and opportunities to designing sustainable AI computing across the key phases of the machine learning (ML) development process — *Data*, *Experimentation*, *Training*, and *Inference* — for a variety of AI use cases at Facebook, such as vision, language, speech, recommendation and ranking. The solution space spans across our fleet of datacenters and on-device computing. Given particular use cases, we consider the impact of AI *data*, *algorithms*, and *system hardware*. Finally, we consider emissions across the life cycle of hardware systems, from manufacturing to operational use.

AI Data Growth. In the past decade, we have seen an exponential increase in AI training data and model capacity. Figure 2(b) illustrates that the amount of training data at Facebook for two recommendation use cases — one of the fastest growing areas of ML usage at Facebook— has increased by 2.4× and 1.9× in the last two years, reaching exabyte scale. The increase in data size has led to a 3.2× increase in data ingestion bandwidth demand. Given this increase, data storage and the ingestion pipeline accounts for a significant portion of

[1] Based on monthly counts, Figure 1 estimates the cumulative number of papers published per category on the arXiv database.

Carbon footprint mapped to the AI lifecycle

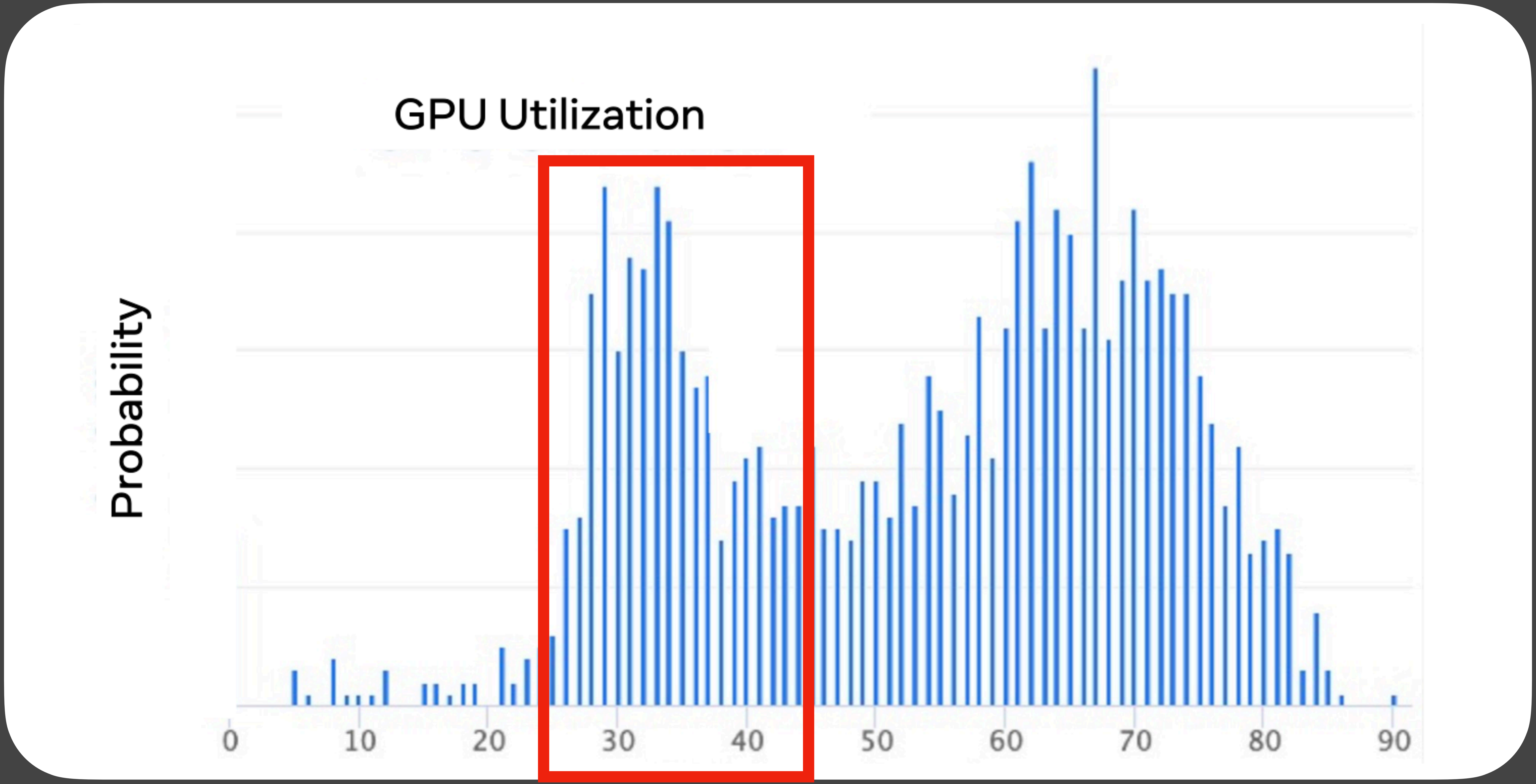


- There are 4 main overarching stages where carbon emissions need to be isolated: **data collection**, **experimentation**, **training**, **inference**.
- At Facebook, recommendation systems split energy consumption **evenly between training and inference**; text translation models have a **35%/65%** split. (Operational cost)
- Operational/embodied cost split: **30%/70%**

Open issues according to Meta

- A vast portion of projects only use **GPUs at 30%**.
  Should be higher to attenuate embodied carbon.

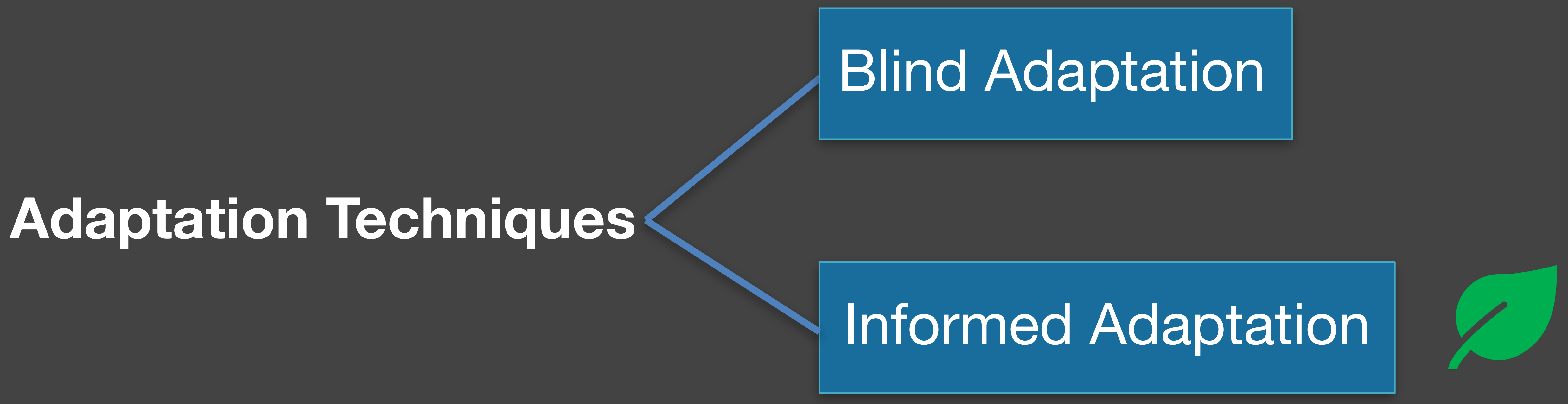
Based on 10K AI projects



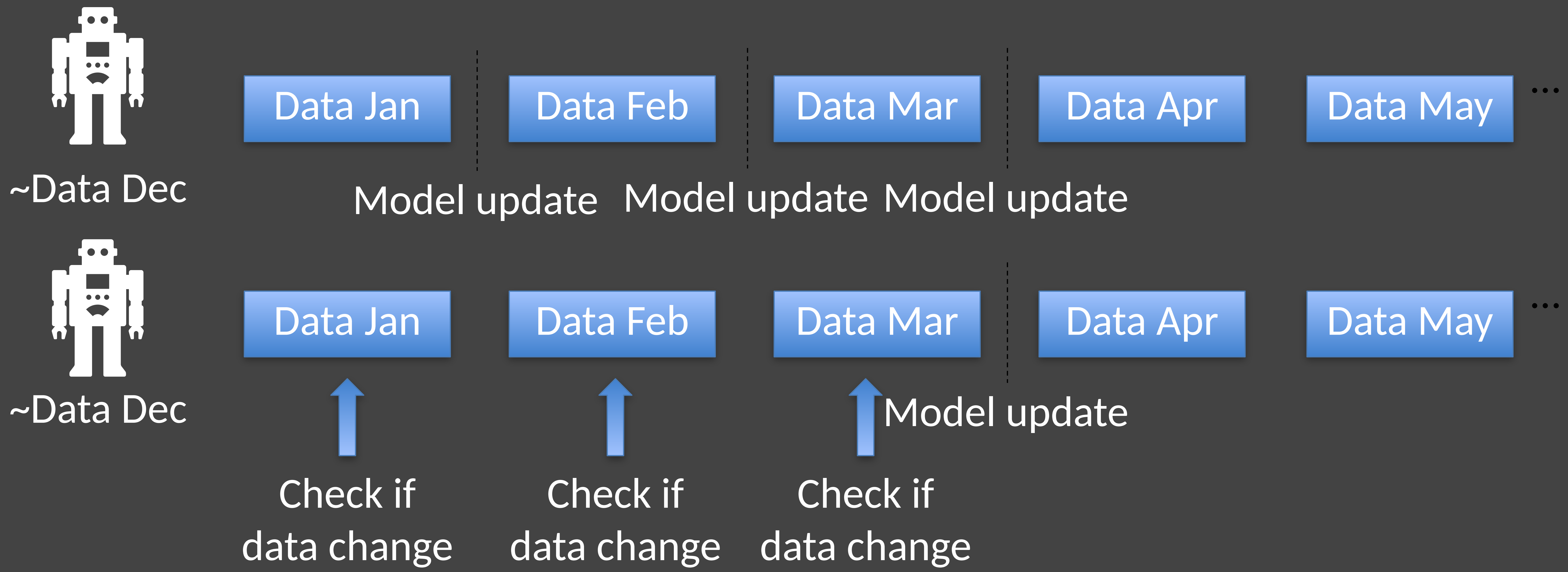
?

Know when to retrain models

Neither too early nor too late



The AI Model will be updated fewer times and only when necessary.



Green Architectural Tactics for ML-Enabled System

ICSE-SEIS 2024

A Synthesis of Green Architectural Tactics for ML-Enabled Systems

Heli Järvenpää
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
h.m.jarvenpaae@student.vu.nl

Patricia Lago
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
p.lago@vu.nl

Justus Bogner
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
j.bogner@vu.nl

Grace Lewis
Carnegie Mellon Software
Engineering Institute
Pittsburgh, PA, USA
glewis@sei.cmu.edu

Henry Muccini
FrAmeLab, University of L'Aquila
L'Aquila, Italy
henry.muccini@univaq.it

Ipek Ozkaya
Carnegie Mellon Software
Engineering Institute
Pittsburgh, PA, USA
ozkaya@sei.cmu.edu

ABSTRACT

The rapid adoption of artificial intelligence (AI) and machine learning (ML) has generated growing interest in understanding their environmental impact and the challenges associated with designing environmentally friendly ML-enabled systems. While Green AI research, i.e., research that tries to minimize the energy footprint of AI, is receiving increasing attention, very few concrete guidelines are available on how ML-enabled systems can be designed to be more environmentally sustainable. In this paper, we provide a catalog of 30 green architectural tactics for ML-enabled systems to fill this gap. An architectural tactic is a high-level design technique to improve software quality, in our case environmental sustainability. We derived the tactics from the analysis of 51 peer-reviewed publications that primarily explore Green AI, and validated them using a focus group approach with three experts. The 30 tactics we identified are aimed to serve as an initial reference guide for further exploration into Green AI from a software engineering perspective, and assist in designing sustainable ML-enabled systems. To enhance transparency and facilitate their widespread use and extension, we make the tactics available online in easily consumable formats. Wide-spread adoption of these tactics has the potential to substantially reduce the societal impact of ML-enabled systems regarding their energy and carbon footprint.

CCS CONCEPTS

• **Software and its engineering** → **Designing software**; **Software architectures**; • **Social and professional topics** → **Sustainability**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Software architecture, architectural tactics, ML-enabled systems, environmental sustainability, Green AI.

ICSE-SEIS '24, April 14–20, 2024, Lisbon, Portugal
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0499-4/24/04.
https://doi.org/10.1145/3639475.3640111

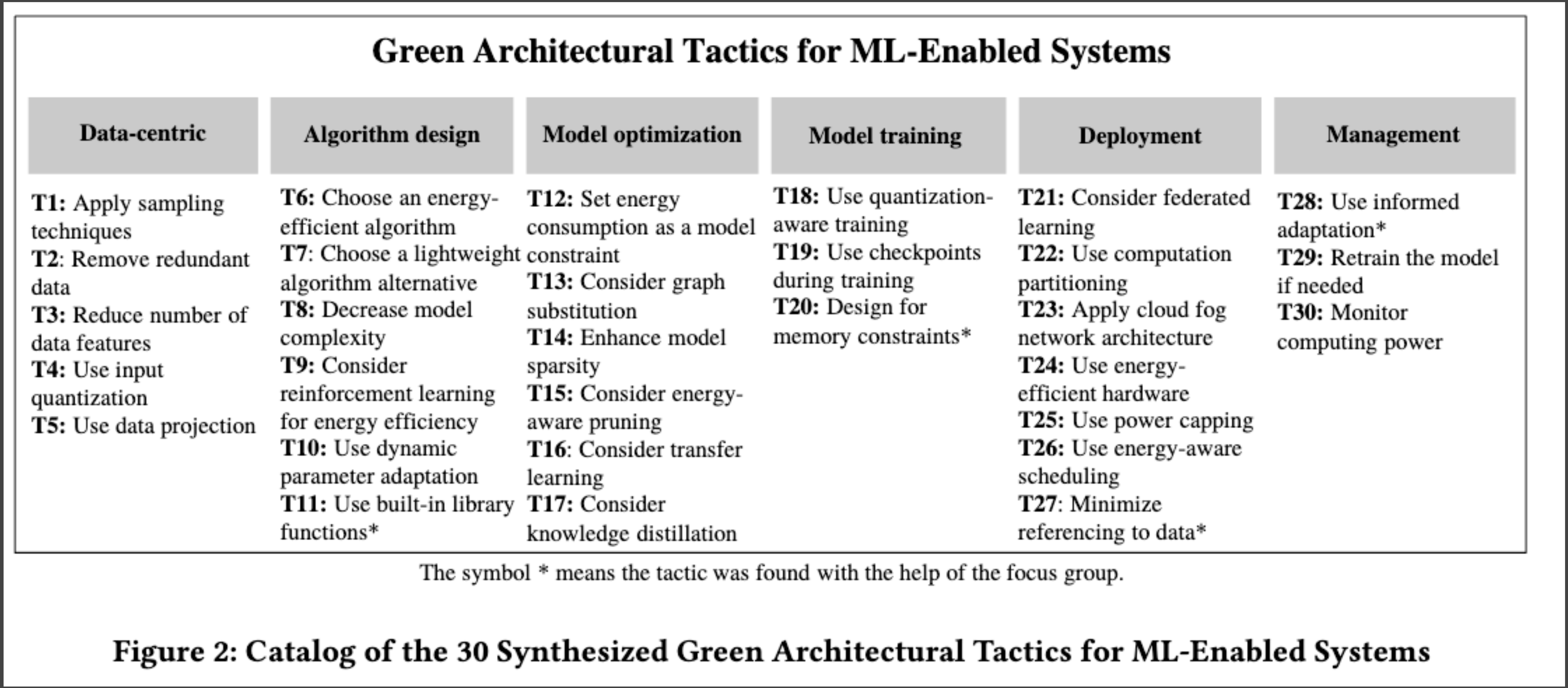
ACM Reference Format:
Heli Järvenpää, Patricia Lago, Justus Bogner, Grace Lewis, Henry Muccini, and Ipek Ozkaya. 2024. A Synthesis of Green Architectural Tactics for ML-Enabled Systems. In *Software Engineering in Society (ICSE-SEIS'24), April 14–20, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3639475.3640111

Lay Abstract: Machine learning (ML) is a technology field that wants to provide software with functionality similar to human-like intelligence, e.g., for understanding text or describing images. However, creating and using systems with ML needs a lot more computing power than non-ML systems, which is bad for the environment. Companies therefore need concrete advice on how they can create ML systems that are environmentally sustainable. In this paper, we provide a catalog of 30 green architectural tactics for these systems. An architectural tactic is a high-level design technique to improve software quality, in our case environmental sustainability. To achieve this, we analyzed 51 scientific papers and later discussed with 3 experts to improve and extend our catalog. If many companies start using these tactics, the energy footprint of systems with ML can be greatly reduced.

1 INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have shown significant potential in digital innovations, with a growing number of different ML applications expanding across a wide spectrum of industries, from healthcare to agriculture and management [41]. This rapid growth of ML applications has also drawn attention to its environmental footprint. Several studies have evaluated the carbon emissions of ML [6, 31]. It is widely acknowledged that training and using ML models at scale is computationally demanding, which leads to greenhouse gas emissions. For example, training a typical transformer-based natural language processing (NLP) model produces greenhouse gas emissions equivalent to five average cars in their lifetime [51]. These considerations have led to new concepts such as *Green AI* and *Sustainability of AI*. Traditional AI has aimed to achieve high accuracy while disregarding energy efficiency, but Green AI emphasizes the environmental footprint of AI and focuses on minimizing computation while still producing accurate results [47]. *Sustainability of AI* refers to the environmental impact of the AI model itself, and highlights the responsible development

Architectural tactics



Green Architectural Tactics for ML-Enabled Systems

Data-centric	Algorithm design	Model optimization	Model training	Deployment	Management
T1: Apply sampling techniques **T2**: Remove redundant data **T3:** Reduce number of data features **T4:** Use input quantization **T5:** Use data projection	**T6:** Choose an energy-efficient algorithm **T7**: Choose a lightweight algorithm alternative **T8:** Decrease model complexity **T9:** Consider reinforcement learning for energy efficiency **T10:** Use dynamic parameter adaptation **T11:** Use built-in library functions*	**T12:** Set energy consumption as a model constraint **T13:** Consider graph substitution **T14:** Enhance model sparsity **T15:** Consider energy-aware pruning **T16**: Consider transfer learning **T17:** Consider knowledge distillation	**T18:** Use quantization-aware training **T19:** Use checkpoints during training **T20:** Design for memory constraints*	**T21:** Consider federated learning **T22:** Use computation partitioning **T23:** Apply cloud fog network architecture **T24:** Use energy-efficient hardware **T25:** Use power capping **T26:** Use energy-aware scheduling **T27**: Minimize referencing to data*	**T28:** Use informed adaptation* **T29:** Retrain the model if needed **T30:** Monitor computing power

The symbol * means the tactic was found with the help of the focus group.

Figure 2: Catalog of the 30 Synthesized Green Architectural Tactics for ML-Enabled Systems

Data-centric

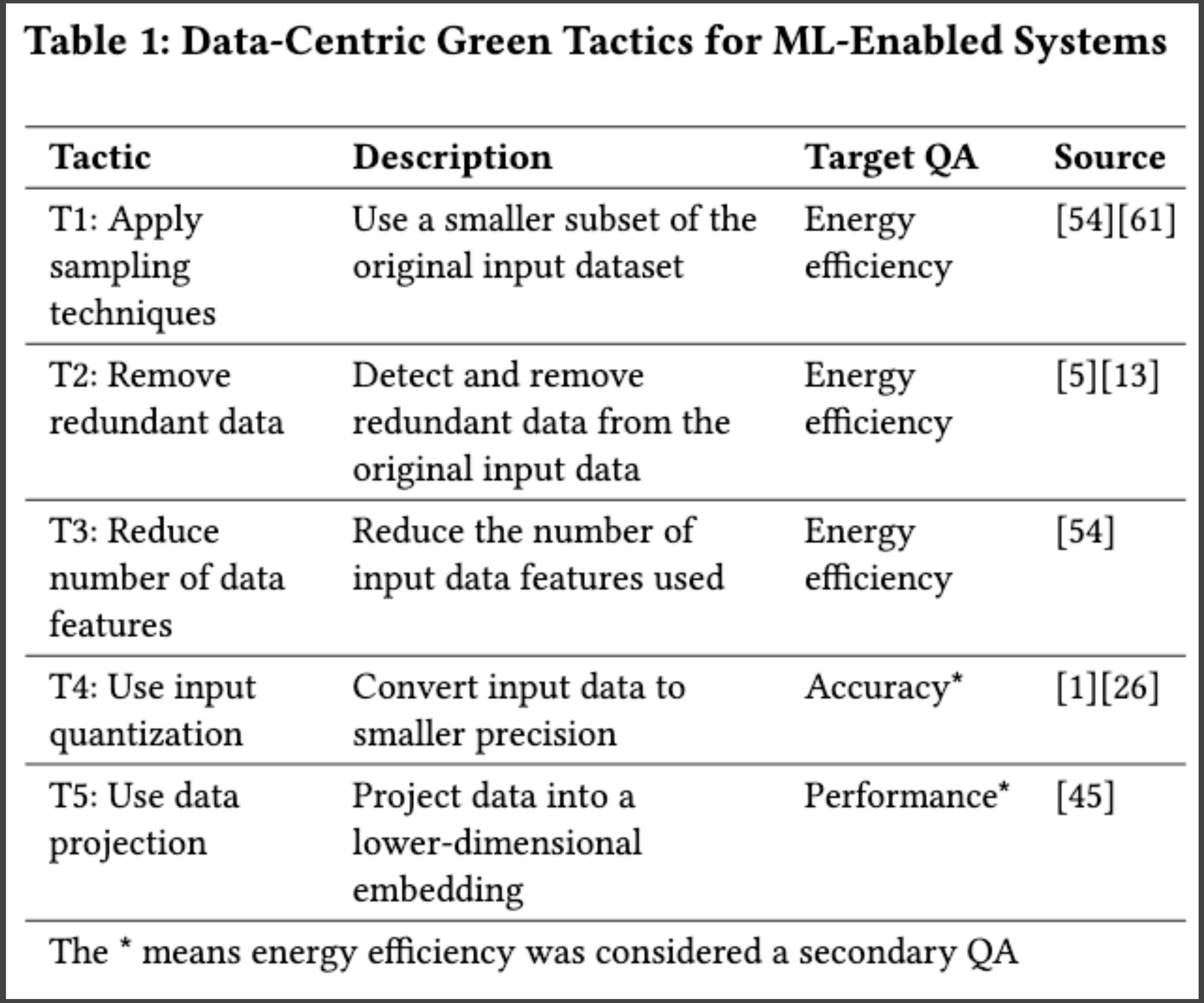


Table 1: Data-Centric Green Tactics for ML-Enabled Systems

Tactic	Description	Target QA	Source
T1: Apply sampling techniques	Use a smaller subset of the original input dataset	Energy efficiency	[54][61]
T2: Remove redundant data	Detect and remove redundant data from the original input data	Energy efficiency	[5][13]
T3: Reduce number of data features	Reduce the number of input data features used	Energy efficiency	[54]
T4: Use input quantization	Convert input data to smaller precision	Accuracy*	[1][26]
T5: Use data projection	Project data into a lower-dimensional embedding	Performance*	[45]

The * means energy efficiency was considered a secondary QA

- Reduce data size
  - Sampling
  - Dimensionality reduction
  - Quantization

Algorithm Design

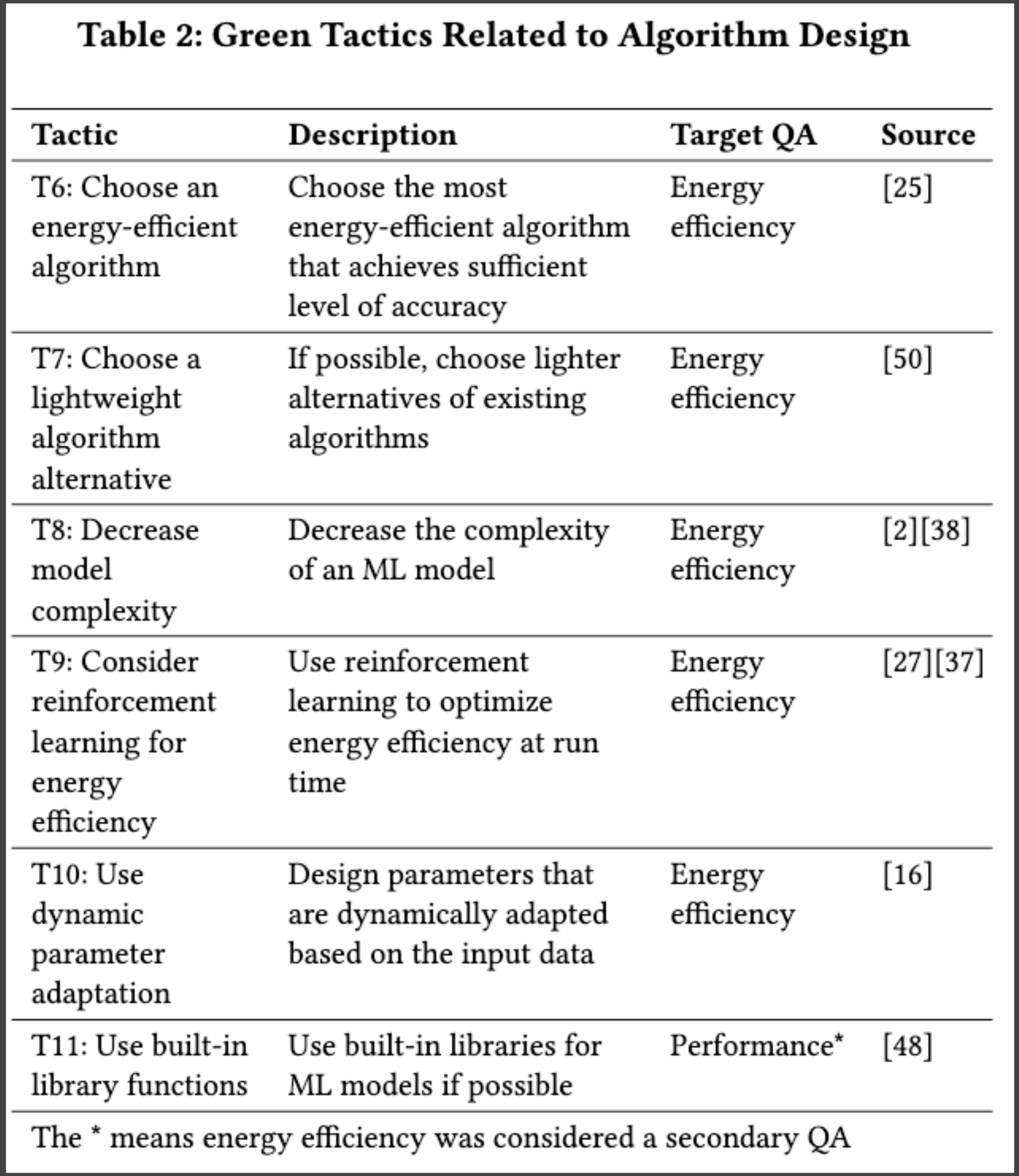


Table 2: Green Tactics Related to Algorithm Design

Tactic	Description	Target QA	Source
T6: Choose an energy-efficient algorithm	Choose the most energy-efficient algorithm that achieves sufficient level of accuracy	Energy efficiency	[25]
T7: Choose a lightweight algorithm alternative	If possible, choose lighter alternatives of existing algorithms	Energy efficiency	[50]
T8: Decrease model complexity	Decrease the complexity of an ML model	Energy efficiency	[2][38]
T9: Consider reinforcement learning for energy efficiency	Use reinforcement learning to optimize energy efficiency at run time	Energy efficiency	[27][37]
T10: Use dynamic parameter adaptation	Design parameters that are dynamically adapted based on the input data	Energy efficiency	[16]
T11: Use built-in library functions	Use built-in libraries for ML models if possible	Performance*	[48]

The * means energy efficiency was considered a secondary QA

- Carefully select your algorithm
- You don’t need the fanciest techniques

Model Optimization

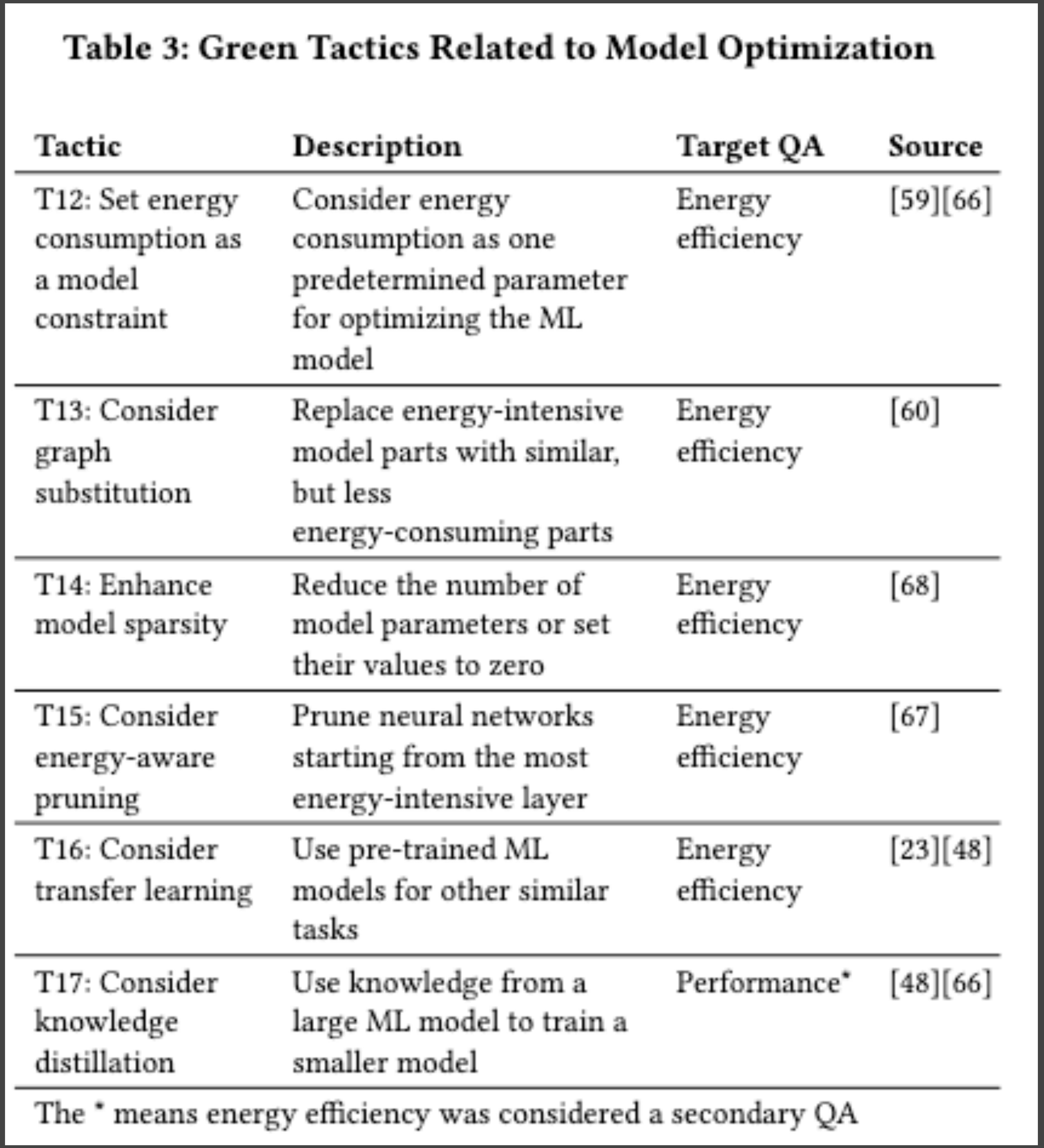


Table 3: Green Tactics Related to Model Optimization

Tactic	Description	Target QA	Source
T12: Set energy consumption as a model constraint	Consider energy consumption as one predetermined parameter for optimizing the ML model	Energy efficiency	[59][66]
T13: Consider graph substitution	Replace energy-intensive model parts with similar, but less energy-consuming parts	Energy efficiency	[60]
T14: Enhance model sparsity	Reduce the number of model parameters or set their values to zero	Energy efficiency	[68]
T15: Consider energy-aware pruning	Prune neural networks starting from the most energy-intensive layer	Energy efficiency	[67]
T16: Consider transfer learning	Use pre-trained ML models for other similar tasks	Energy efficiency	[23][48]
T17: Consider knowledge distillation	Use knowledge from a large ML model to train a smaller model	Performance*	[48][66]

The * means energy efficiency was considered a secondary QA

- Add energy to training parameters
- Reduce FLOPs
  - Pruning, sparsity
- Take advantage of existing models

Model Training

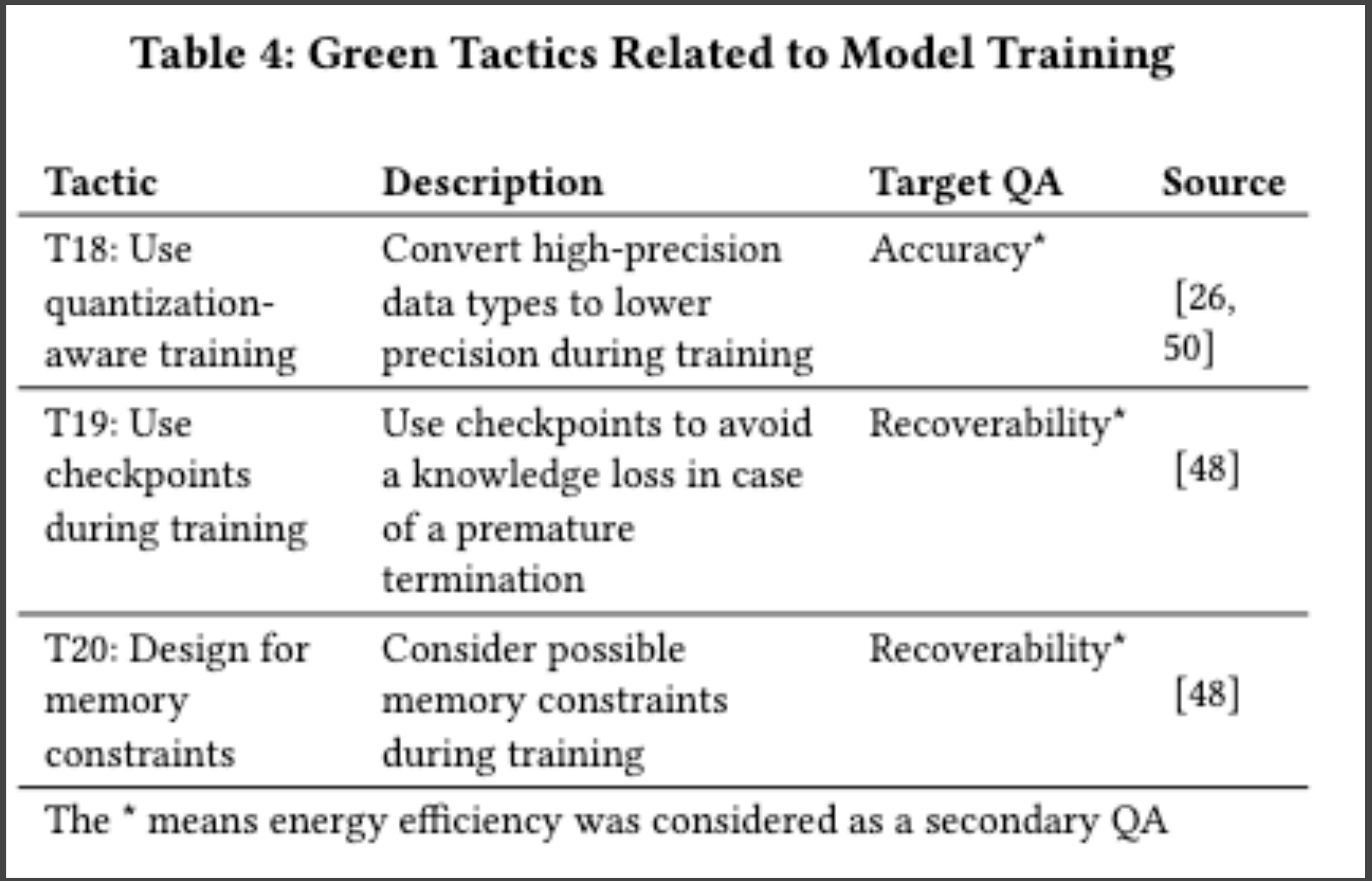


Table 4: Green Tactics Related to Model Training

Tactic	Description	Target QA	Source
T18: Use quantization-aware training	Convert high-precision data types to lower precision during training	Accuracy*	[26, 50]
T19: Use checkpoints during training	Use checkpoints to avoid a knowledge loss in case of a premature termination	Recoverability*	[48]
T20: Design for memory constraints	Consider possible memory constraints during training	Recoverability*	[48]

The * means energy efficiency was considered as a secondary QA

- Quantization
- SAVE TRAINING PROGRESS

Model Deployment

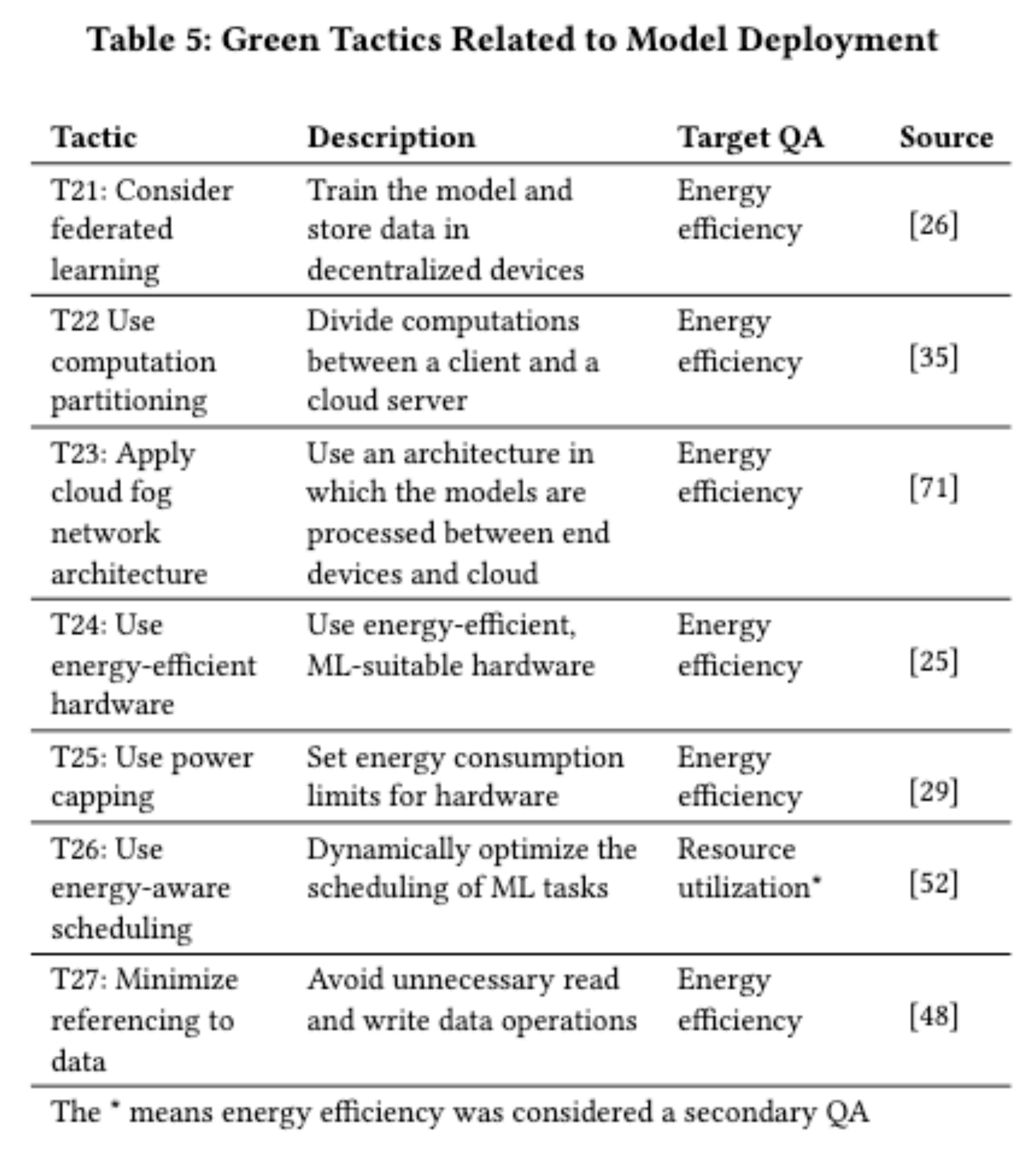


Table 5: Green Tactics Related to Model Deployment

Tactic	Description	Target QA	Source
T21: Consider federated learning	Train the model and store data in decentralized devices	Energy efficiency	[26]
T22 Use computation partitioning	Divide computations between a client and a cloud server	Energy efficiency	[35]
T23: Apply cloud fog network architecture	Use an architecture in which the models are processed between end devices and cloud	Energy efficiency	[71]
T24: Use energy-efficient hardware	Use energy-efficient, ML-suitable hardware	Energy efficiency	[25]
T25: Use power capping	Set energy consumption limits for hardware	Energy efficiency	[29]
T26: Use energy-aware scheduling	Dynamically optimize the scheduling of ML tasks	Resource utilization*	[52]
T27: Minimize referencing to data	Avoid unnecessary read and write data operations	Energy efficiency	[48]

The * means energy efficiency was considered a secondary QA

- Distributed deployment
- Energy efficient hardware and configurations

Model Management

- Reuse the model as much as possible

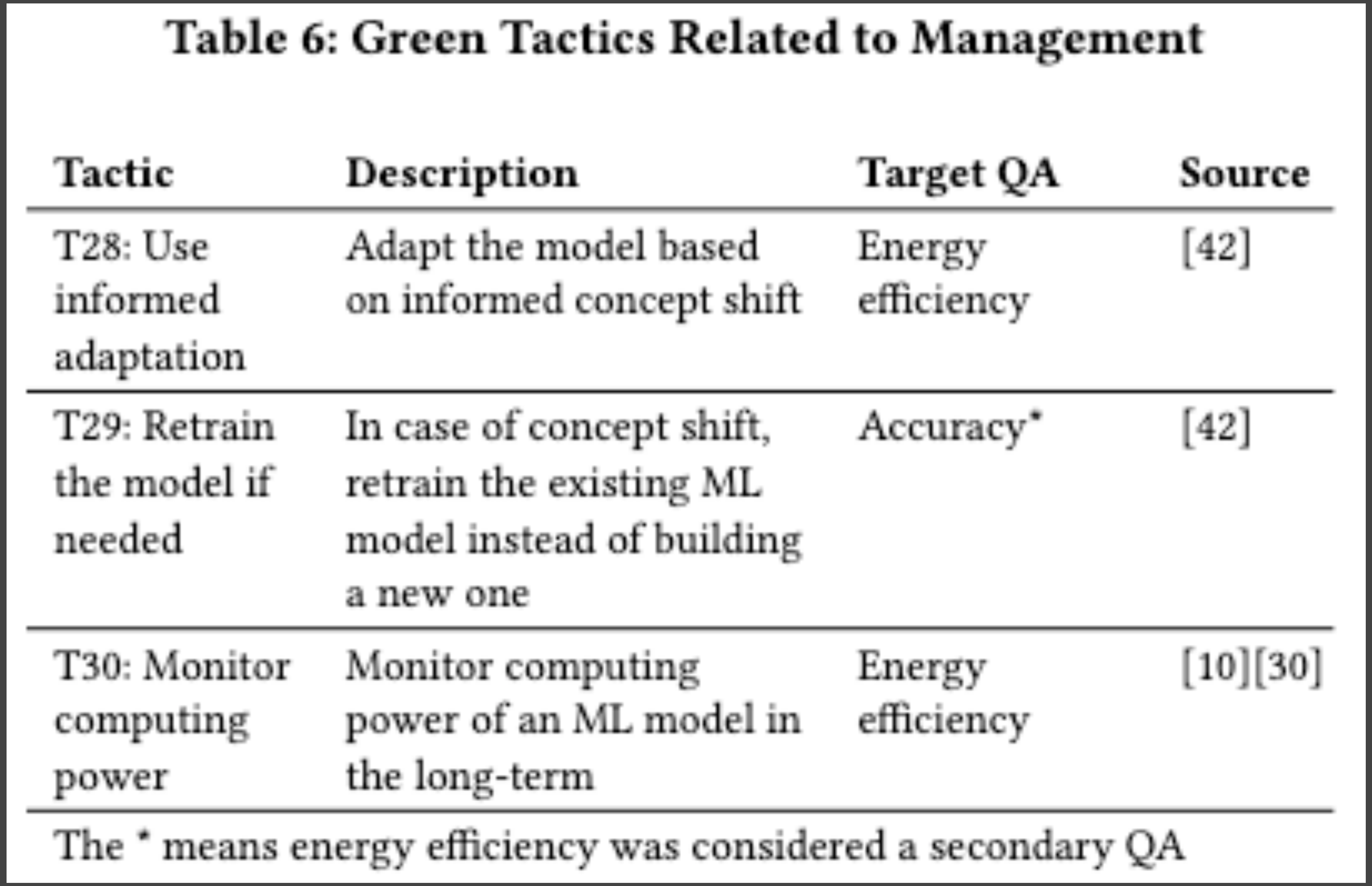
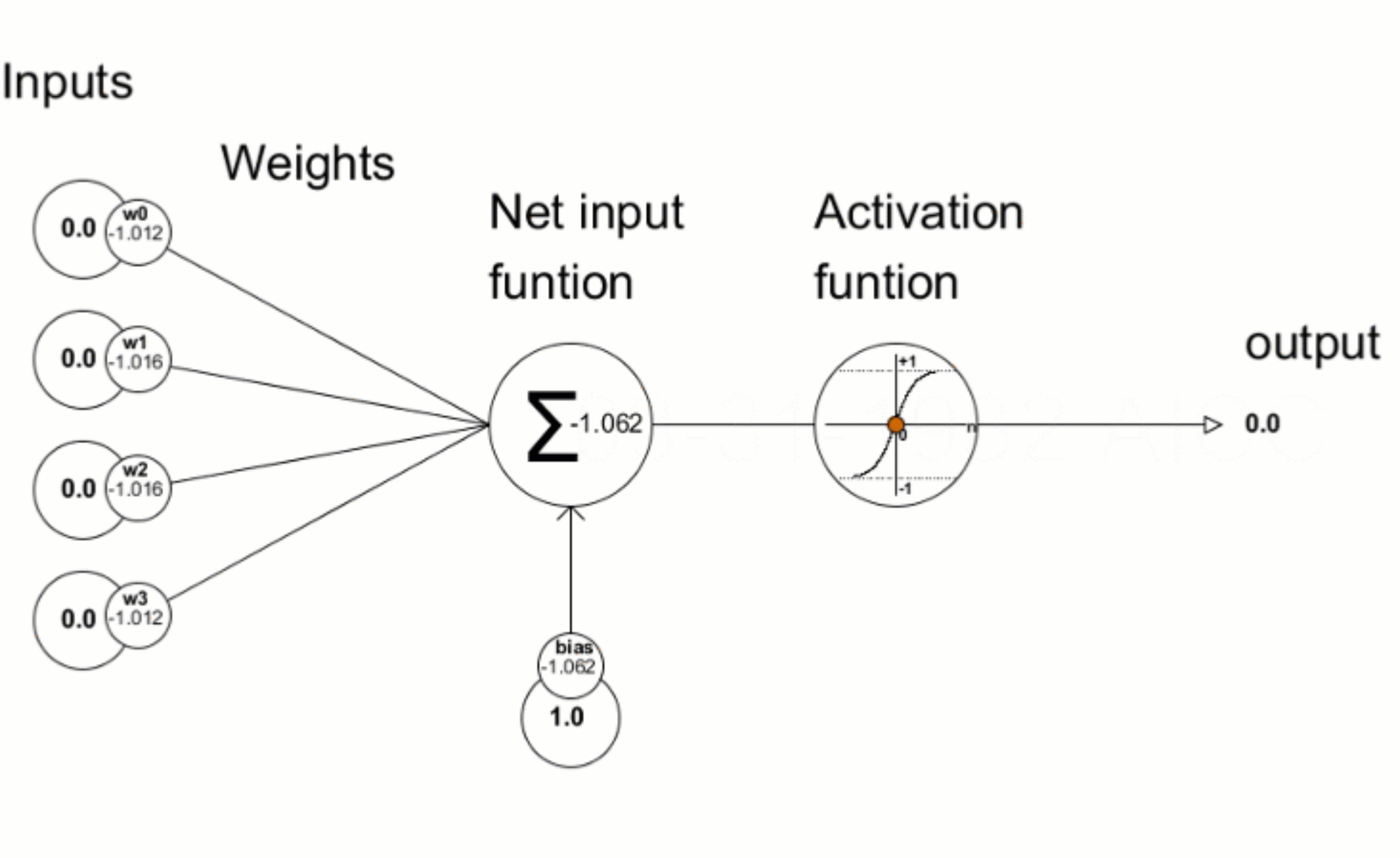


Table 6: Green Tactics Related to Management

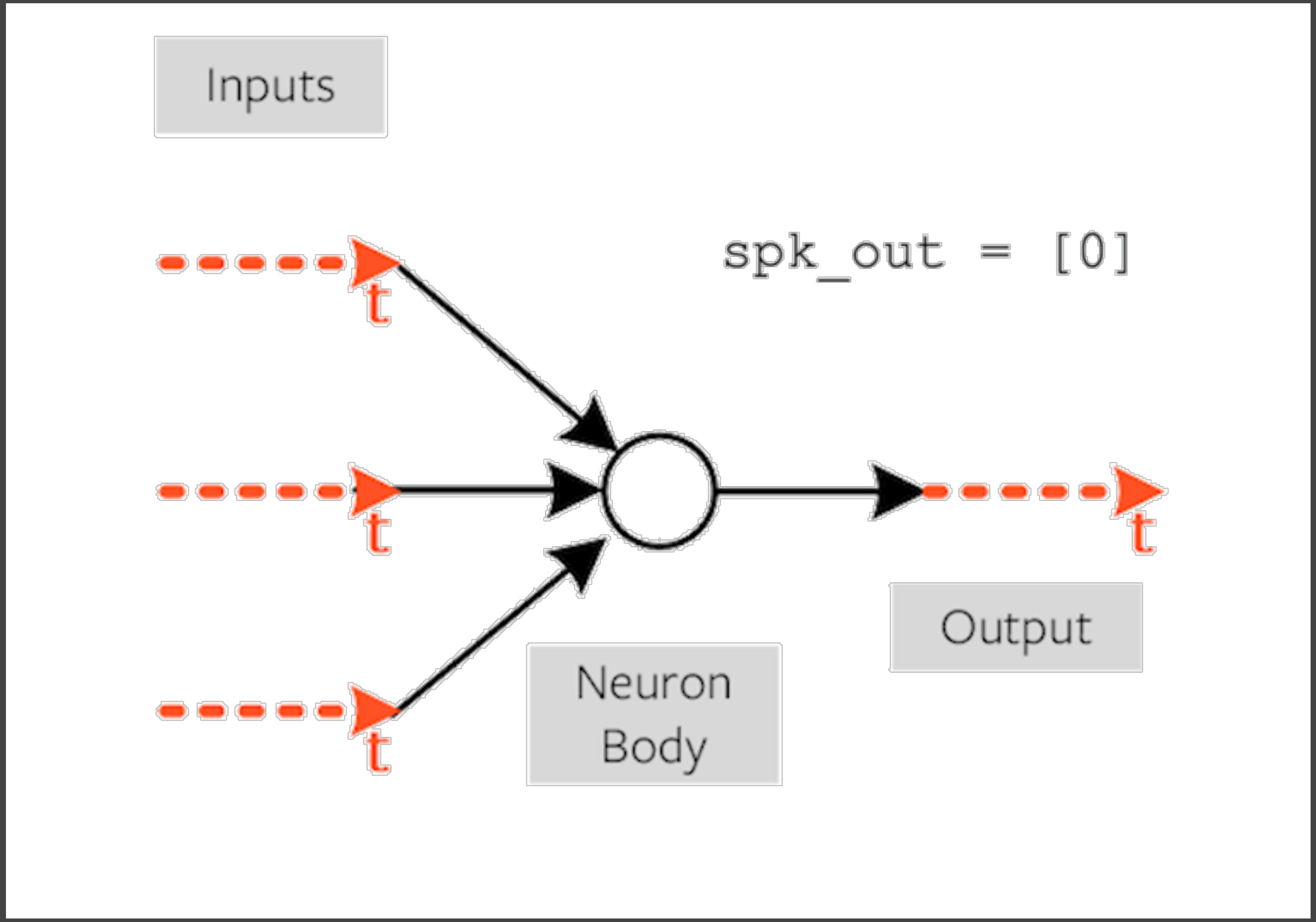
Tactic	Description	Target QA	Source
T28: Use informed adaptation	Adapt the model based on informed concept shift	Energy efficiency	[42]
T29: Retrain the model if needed	In case of concept shift, retrain the existing ML model instead of building a new one	Accuracy*	[42]
T30: Monitor computing power	Monitor computing power of an ML model in the long-term	Energy efficiency	[10][30]

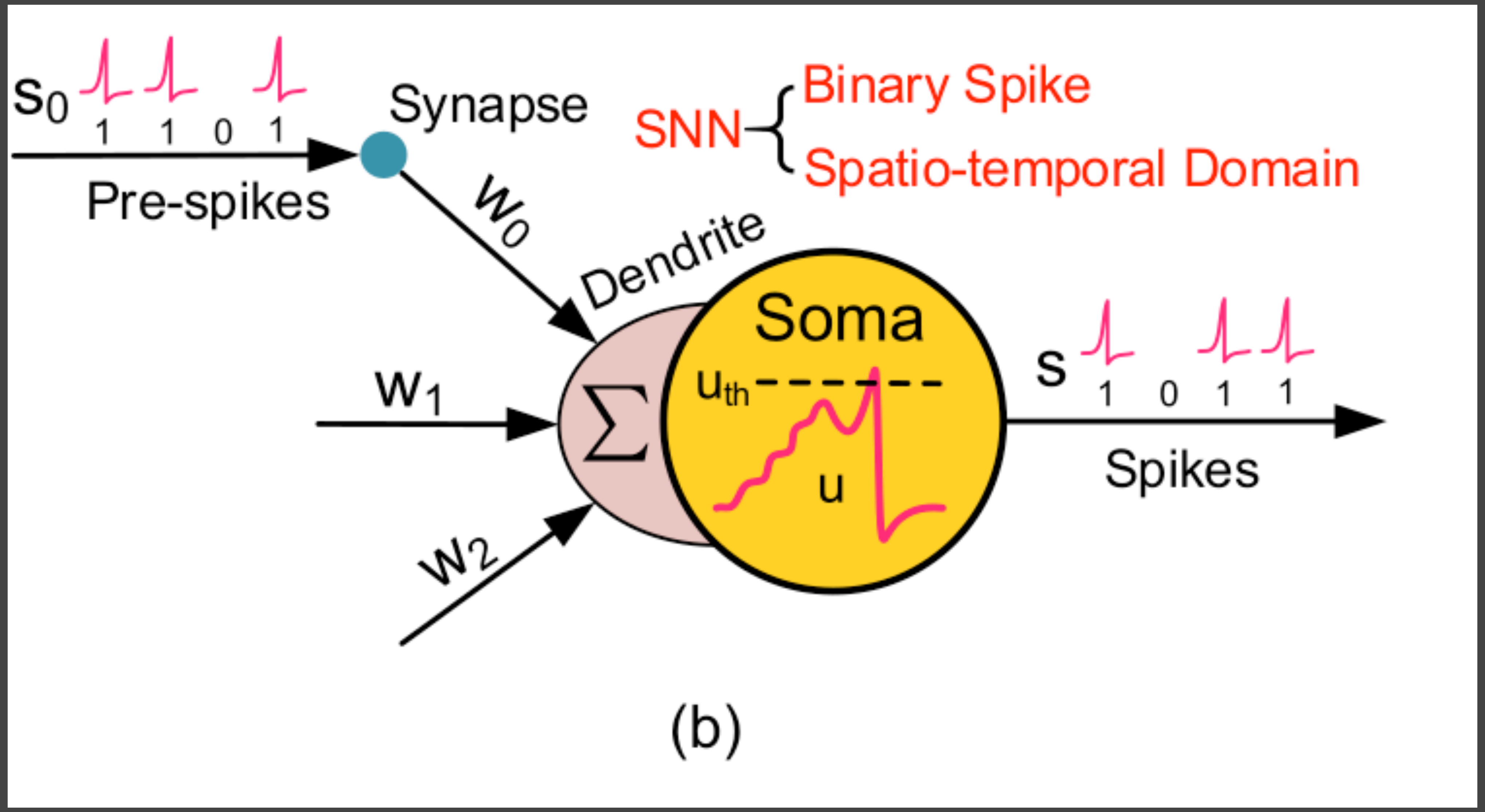
The * means energy efficiency was considered a secondary QA

Rethinking the Architecture: Spiking Neural Networks

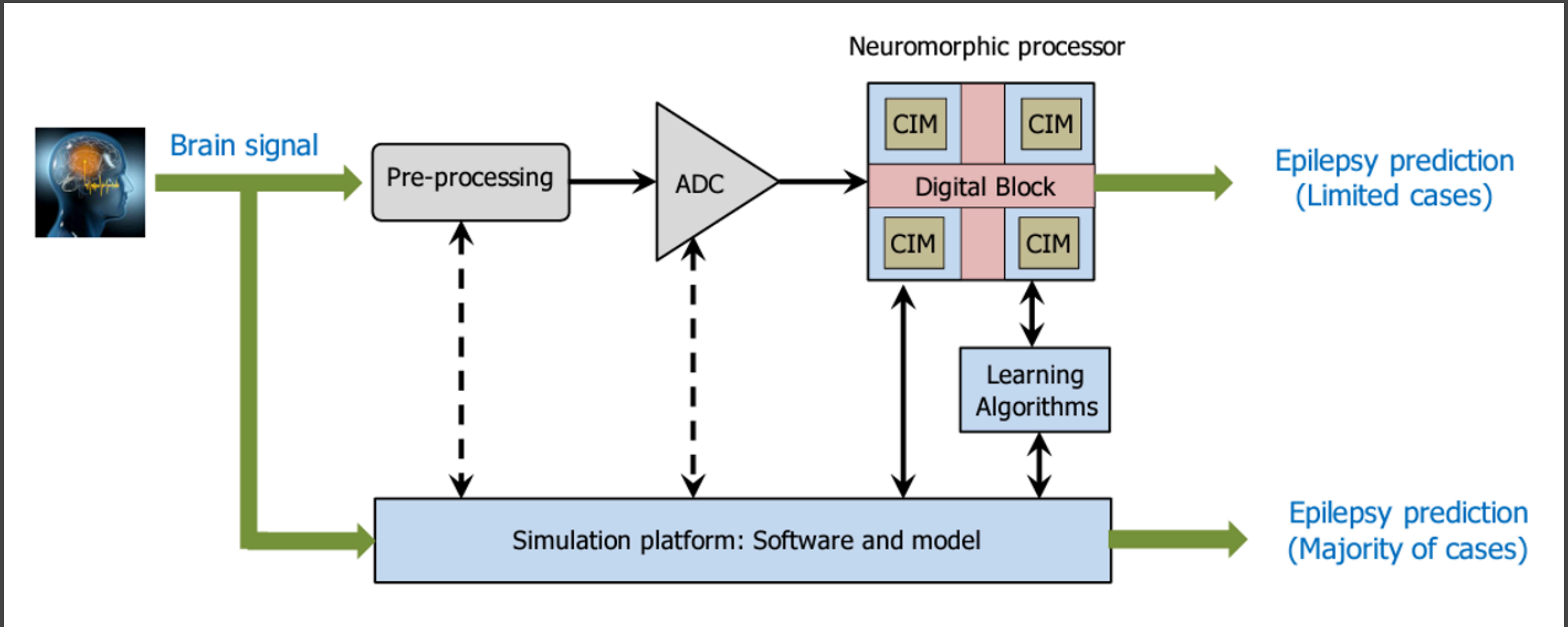


Rethinking the Architecture: Spiking Neural Networks





SELF Lab


Neuromorphic processor
Brain signal
Pre-processing
ADC
CIM
CIM
Digital Block
CIM
CIM
Epilepsy prediction
(Limited cases)
Learning
Algorithms
Simulation platform: Software and model
Epilepsy prediction
(Majority of cases)

recap