UNIT III

Green Software Engineering and Sustainable Artificial Intelligence

The core AI/DS-specific unit: how algorithm and system design choices — not just hardware — determine energy and carbon cost.

06 Hours

Dr. Sneha Thombre
MKSSS's Cummins *College* of Engineering for Women, Pune

Faculty Orientation Deck · Open Elective — Third Year AI & Data Science · Savitribai Phule Pune University

ROADMAP

Unit at a Glance

Purpose: The strongest justification for offering this course to AI & Data Science students — it addresses their own workflows directly: model training, coding practices, and cloud usage.

- Energy consumption of software systems
- Green software engineering principles: efficient algorithms, optimised coding, sustainable architecture
- Green AI techniques: model compression, pruning, efficient training
- Energy-aware data science workflows; carbon-aware computing
- Scope 1/2/3 GHG accounting for technology organisations
- Emission calculation from activity data (electricity, cloud usage, transport)
- Generative AI models and their environmental impact
- AI-driven analytics for sustainability monitoring & optimisation
- Carbon accounting tools, ESG analytics platforms, sustainability dashboards

Maps chiefly to CO4 (Implement energy-efficient techniques in AI systems) · Case study: Energy-Efficient ML Algorithms & Sustainable AI Model Design

CORE CONCEPT

Green Software Engineering: Where Energy Is Actually Spent

Inefficient code isn't just slow — it draws more electricity for the same task. Four levers students should learn:

1 Efficient Algorithms

Lower time/space complexity directly reduces compute cycles and therefore energy — an O(n log n) vs $O(n^2)$ choice is a sustainability choice.

2 Optimised Coding

Avoiding redundant computation, unnecessary I/O, and excess memory allocation cuts energy per request or per training step.

3 Sustainable Architecture

Caching, batching, and right-sizing infrastructure to actual load avoids idle over-provisioning of servers and GPUs.

4 Carbon-Aware Scheduling

Running non-urgent batch/training jobs when the grid's electricity mix is cleaner (more solar/wind) reduces the same job's emissions.

CORE CONCEPT

Green AI: Making Models Efficient, Not Just Accurate

- "Green AI" (Schwartz et al.) argues efficiency — compute, data, energy — should be reported and optimised for, alongside accuracy, not treated as an afterthought.
- Model compression: shrinking a trained model's size while preserving most of its accuracy, reducing storage and inference energy.
  - Pruning: removing redundant weights/neurons that contribute little to output — smaller, faster, cheaper to run.
  - Quantization: using lower-precision number formats (e.g. 8-bit instead of 32-bit) for weights, cutting memory and compute cost.
- Efficient training: techniques like early stopping, transfer learning, and smaller-but-well-chosen datasets reduce training energy without gutting performance.
- Energy-aware data science workflows apply these choices throughout a pipeline — not just at the final model — including data preprocessing and experimentation.

60%

Share of AI energy use estimated to go toward inference (using the model), not training (Google estimate)

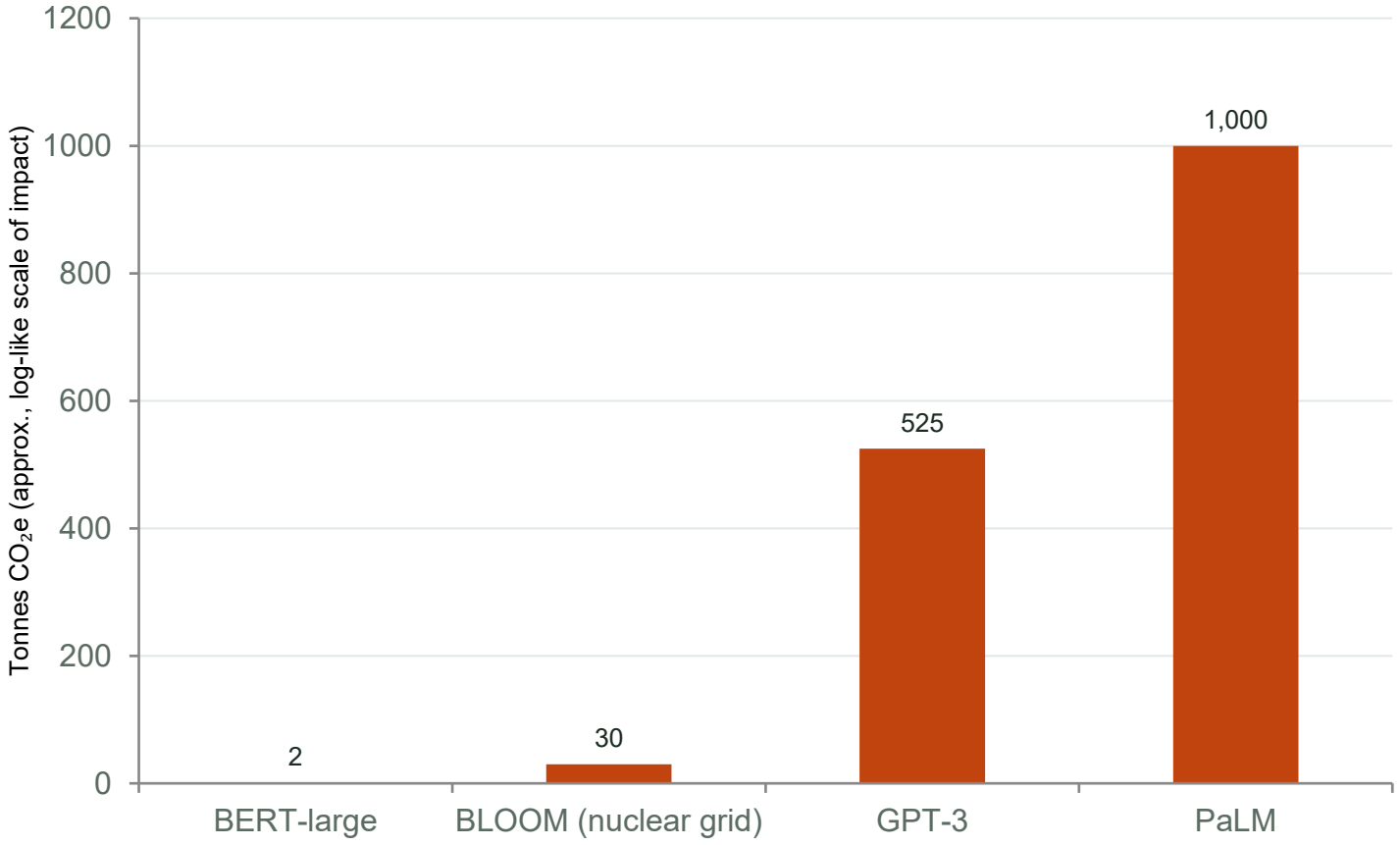
12×

GPT-4's estimated training emissions vs GPT-3, reflecting the trend toward larger models

estimate (2024)

DATA & STATISTICS

Model Size and Training Location Both Drive Emissions



Why this matters

BLOOM (176B parameters) emitted far less than GPT-3 (175B) largely because it trained on a low-carbon (nuclear) grid — proving that where and how you train matters as much as model size. This directly motivates carbon-aware scheduling and Green AI practice.

Sources: Luccioni et al. (2022) BLOOM carbon report; Patterson et al. (2021); multiple secondary estimates, figures approximate and vary by methodology

CASE STUDY (SYLLABUS-PRESCRIBED)

Case Study: Energy-Efficient Machine Learning Algorithms and Sustainable AI Model Design

- A data science team needs a text-classification model. Option A: fine-tune a 7-billion-parameter model from scratch. Option B: apply transfer learning on a pre-trained, smaller model and add pruning/quantization for deployment.
- Option B typically achieves comparable accuracy on many real tasks at a fraction of the training compute and a much smaller memory/inference footprint — illustrating the Green AI principle of reporting efficiency alongside accuracy.
- At inference time, a quantized (8-bit) version of a deployed model can serve significantly more requests per unit of energy than its full-precision counterpart, with only marginal accuracy loss for most applications.
- Carbon-aware scheduling: the same training job, delayed by a few hours to run when the regional grid draws more from solar/wind, can be reported with a measurably lower Scope 2 footprint for identical compute.

Discuss in class

- When is it justified to accept a small accuracy drop in exchange for a large efficiency gain — and when is it not?
- How would you convince a project team to add an 'emissions per training run' metric to their model reports?
- What's the difference between reducing a model's training footprint and reducing its inference footprint — which matters more for a widely deployed product?
- Could carbon-aware scheduling alone meaningfully cut a company's AI emissions without any model changes?

PEDAGOGY — FOR FACULTY

Suggested Pedagogy for Unit III

1 Live complexity-vs-energy demo

Run two versions of the same task (e.g. an $O(n^2)$ vs $O(n \log n)$ sort, or a brute-force vs vectorised computation) and show the measurable time/power difference on a laptop wattmeter or software estimator.

2 Hands-on: quantize/prune a small model

Use a lightweight framework (e.g. a small scikit-learn or PyTorch model) so students can measure accuracy vs size/latency trade-offs after pruning or quantization themselves.

3 Tool walkthrough: ML CO2 estimation

Demonstrate an open carbon-tracking tool (e.g. CodeCarbon or the ML CO2 Impact calculator) so students can estimate emissions for their own small training runs.

4 Debate: accuracy-first vs efficiency-first AI

Structured debate on whether leaderboards should mandate reporting energy/emissions alongside accuracy — connects directly to the Green AI paper's argument.

5 Dashboard walkthrough

Show a real or simulated ESG/sustainability dashboard (screenshots or a demo) so students see how carbon accounting tools are actually used in industry.

6 Cross-branch relevance bridge

For non-AI/DS students in the open elective, reframe 'green software' more broadly — efficient code and cloud usage apply to any software project, not only ML.

Faculty tip: *This is the most technically hands-on unit — even a 20-minute live coding/measurement demo will do more for retention than any amount of slide-based explanation of pruning or quantization.*

Evaluation Techniques for Unit III

Assessment Component	Weight	How it maps to this unit
CCE quiz on Green AI terminology	**5 Marks**	Define pruning, quantization, model compression, carbon-aware computing, and distinguish training vs inference energy.
Hands-on lab / mini-assignment	**10 Marks**	Apply one efficiency technique (e.g. quantization) to a small model and report the accuracy-vs-efficiency trade-off observed.
Case-study analysis (written or oral)	**5 Marks**	Written or oral reasoning on the GPT-3/BLOOM-style case — tests conceptual application over memorised numbers.
End-Semester theory question	**Included in 35 Marks**	Typically a descriptive/numerical question on Green AI techniques or Scope-based emission calculation for a software system.

Where lab infrastructure allows, prioritise the hands-on assignment over a pure quiz — this unit is the strongest opportunity in the whole syllabus for applied, portfolio-worthy student work.

KEY TAKEAWAYS

Unit III — Key Takeaways for Faculty

- ✓ Software design choices — algorithms, architecture, scheduling — are sustainability choices, not just performance choices.
- ✓ Green AI treats efficiency as a first-class metric alongside accuracy, not an afterthought.
- ✓ Model compression, pruning, and quantization are concrete, teachable, and directly applicable techniques for AI/DS students.
- ✓ Where and when a model is trained (grid carbon intensity, scheduling) can matter as much as model size — a genuinely counter-intuitive and memorable point.

For faculty-led discussion

- Should journals and conferences require an energy/emissions disclosure for published ML models, similar to a data availability statement?
- Is 'bigger models, but trained on a cleaner grid' a real solution, or does it just relocate the problem?
- How should a student weigh Green AI principles against a placement interview that only tests raw model accuracy?