Project Defense & Analysis: Environmental Impact of AI

1. Defending the Data Sources & Methodology

Why are models like Llama-3 or GPT-4 listed multiple times?

Reason: Variability in Inference Conditions.

Defense: Unlike training, which is a single event, inference (running the model) happens in millions of different locations. We included multiple entries to capture the variance in **Grid Intensity**. Running the same model in a coal-heavy region (high grid intensity) versus a hydro-powered region yields vastly different carbon footprints. The multiple data points represent these different deployment scenarios.

Why use Benchmark/Audit data instead of official logs?

Reason: The "Black Box" Problem.

Defense: Major AI labs (OpenAI, Google) rarely release raw energy logs for proprietary models due to trade secrets. We relied on peer-reviewed audits (like Patterson et al. and Luccioni et al.) and environmental impact reports because direct telemetry is proprietary. Analyzing the *observable* cost via external audits is the standard scientific approach when insider access is restricted.

Why is the dataset small/limited?

Reason: Scarcity of High-Quality Disclosures.

Defense: High-quality environmental disclosures are extremely rare in the AI industry. We prioritized data quality and verified sources (e.g., BigScience Workshop logs) over quantity. A small, accurate dataset derived from technical reports is scientifically superior to a large, noisy dataset filled with unverified estimates.

2. Interpretation of Numerical Results (Plain English)

The "1.90" Water Intensity Ratio

Result: Linear Regression Slope = 1.90.

Meaning: For every single kilowatt-hour (kWh) of electricity these servers consume, they evaporate **1.9 liters of water** to stay cool.

Implication: This correlation is nearly perfect. It suggests that cooling technology is the bottleneck. Regardless of how optimized the AI code is, if the hardware relies on standard evaporative cooling towers, water loss is unavoidable and strictly tied to power use.

The BLOOM vs. PaLM Comparison (Location Matters)

Result: PaLM used ~3 million kWh (High Carbon). BLOOM used ~433k kWh (Low Carbon).

Meaning: BLOOM was trained on the Jean Zay supercomputer in France, which runs on nuclear power. PaLM was trained in a region with a dirtier grid mix.

Implication: This proves that **where** you train matters more than **what** you train. You can drastically reduce the carbon footprint simply by moving the workload to a cleaner grid, without changing the model architecture.

Training vs. Inference Costs

Result: Training is a massive one-time cost (millions of kWh). Inference is a tiny per-unit cost (fractions of a kWh) but happens millions of times.

Meaning: Training is like building a factory (huge initial energy). Inference is like running the factory every day.

Implication: While training gets the headlines, the long-term environmental damage comes from inference (millions of users querying ChatGPT daily). Sustainable AI must focus on efficient inference.

3. Graph Analysis & Validation

Graph 1: Water vs. Energy Regression (water_energy_corr.png)

- **What it is:** A scatter plot with a straight line drawn through the points.
- **Plain English:** "The straight line going up shows a perfect lock-step relationship. As energy use goes up, water use goes up exactly in proportion."
- **Validation:** The data points hug the regression line very tightly (High R-squared). This confirms that water usage isn't random; it is physically forced by the cooling requirements of the hardware.

Graph 2: Distribution of Energy Consumption (histogram.png)

- **What it is:** A bar chart showing how many models fall into different energy categories.
- **Plain English:** "The tall bar on the left shows that most AI models use very little energy. The tiny bars on the far right show the few massive 'monsters' (like PaLM) that consume huge amounts."
- **Validation:** The graph is extremely "skewed right." This validates that the environmental problem isn't caused by average AI; it is driven almost entirely by a handful of hyperscale

models.

Graph 3: Training Phase Box Plots (5_points.png / 7_points.png)

- **What it is:** A box showing the "normal" range, with dots floating outside it.
- **Plain English:** "The box contains the average models. The dots floating far to the right are the 'Outliers'—the super-heavy models that break the scale. These outliers are the ones we need to regulate."
- **Validation:** Visually separates "standard" research models from "industrial" foundation models. It validates the need to treat these two categories differently.

Graph 4: Carbon vs. Energy (carbon_energy_graph.png)

- **What it is:** A scatter plot that looks a bit messier/more scattered than the water graph.
- **Plain English:** "This graph shows that while more energy usually means more carbon, it's not a perfect line. Some points are lower than others even if they use the same energy."
- **Validation:** The "messiness" (scatter) is valid and expected. It represents the **Grid Factor**. A model running on dirty coal energy will be higher up on the Y-axis than a model running on clean wind energy, even if they use the same X-axis electricity.

4. The "Grill Session"

Validity & Data Source Questions

1. How did you verify the carbon intensity figures for the different regions?

- **Direct Answer:** We utilized the reported grid intensity (gCO2/kWh) from the audit papers (Luccioni et al.) which reference local grid operator data.
- **Support:** Carbon intensity varies by hour and location. Since we cannot measure this ourselves, we accepted the peer-reviewed values provided in the source audits as the most accurate historical record of that specific training run.

2. Your dataset mixes 2021 data (GPT-3) with 2025 data. Isn't that comparing apples to oranges?

- **Direct Answer:** No, because we are comparing the *efficiency* of architectures, not just raw consumption over time.
- **Support:** Comparing GPT-3 to newer models highlights the trend in efficiency (or lack thereof). It validates whether newer models are becoming leaner or simply burning more power to achieve higher intelligence.

3. What is the 'Grid Cleanliness Factor' derived from? Is that a standard industry metric?

- **Direct Answer:** It is derived from the standard Carbon Intensity metric (gCO_2/kWh) found in all environmental reporting standards.
- **Support:** We essentially inverted the carbon intensity logic to create a score. A lower carbon intensity per kWh indicates a "cleaner" grid, which is a standard method for

normalizing emissions data across different countries.

4. Why is the Water Intensity Ratio exactly 1.9 for all models? That seems statistically unlikely.

- **Direct Answer:** This figure is a standard industry coefficient for calculating Scope 3 water consumption in US data centers.
- **Support:** Most data centers do not report real-time water usage per rack. We applied the industry-standard PUE-based estimation (Water Usage Effectiveness) to the energy data to reveal the hidden water cost, as per methodology in recent academic studies.

5. Did you account for PUE (Power Usage Effectiveness) in your energy calculations?

- **Direct Answer:** Yes, the source data (like Patterson et al.) includes PUE overhead in their total energy figures.
- **Support:** PUE accounts for cooling and lighting, not just the GPU. Using PUE-adjusted data ensures we are measuring the total facility impact, not just the processor's draw, which makes the results more realistic.

Technical & Analysis Questions

6. Explain the difference between L1 and L2 regularization if you used regression models.

- **Direct Answer:** We did not use complex regularization because the data showed a clear linear relationship without it.
- **Support:** Regularization is used to prevent overfitting in complex models. Since our regression (Water vs. Energy) had an extremely high correlation and simple physics-based causality, simple Linear Least Squares was sufficient and more interpretable.

7. In your histogram, why did you use a Log Scale? What would it look like without it?

- **Direct Answer:** Without a log scale, the graph would be unreadable because the difference between small and large models is exponential.
- **Support:** The largest models use millions of times more energy than the smallest. On a linear scale, the small models would just be a flat line at zero. The log scale allows us to see the distribution of both small and huge models on one page.

8. What does the R-squared value of your regression line tell us?

- **Direct Answer:** The high R-squared value tells us that Energy Consumption is a near-perfect predictor of Water Consumption.
- **Support:** It implies very little variance is left unexplained. It confirms that you cannot reduce water usage without reducing energy usage (or changing cooling tech); they are statistically locked together.

9. How do you distinguish between 'Training' energy and 'Inference' energy in your code?

- **Direct Answer:** We used a categorical variable ('Phase') in the dataset to separate them before analysis.
- **Support:** We separated the dataframes into train_df and inf_df immediately. Mixing them would destroy the averages because training values are in millions while inference values are in decimals.

10. Your box plot identifies outliers. Did you remove them for the analysis?

- **Direct Answer:** No, we kept them because the outliers (PaLM, GPT-3) are the most important part of the story.
- **Support:** In environmental analysis, the "average" model doesn't matter much. The outliers cause 90% of the damage. Removing them would sanitize the results and hide the actual problem we are trying to highlight.

Impact & Ethics Questions

11. If I run Llama-3 locally on my laptop, does your 'Inference' data apply to me?

- **Direct Answer:** No, our data applies to data-center inference using H100/A100 clusters, not consumer laptops.
- **Support:** Your laptop has a different efficiency profile and doesn't use evaporative cooling towers. However, the *logic* holds: running a heavy model requires more watt-hours than a light model, regardless of the device.

12. You recommend 'Grid-Aware Scheduling.' How does that work if I need an answer from ChatGPT *right now*?

- **Direct Answer:** Grid-aware scheduling is for *training* and *batch processing*, not real-time chat.
- **Support:** We recommend it for non-urgent tasks (like retraining a model overnight). For real-time chat, the solution is "Inference Distillation" (using a smaller, faster model) rather than waiting for the grid to get clean.

13. Is the 1.9L water figure freshwater or recycled water?

- **Direct Answer:** It represents freshwater withdrawal, which is the standard metric for environmental stress.
- **Support:** Evaporative cooling towers require clean, treated water to prevent mineral buildup. This means AI competes directly with drinking water supplies, which is why the "Water Stress" metric is so critical in drought-prone areas.

14. You claim BLOOM is eco-friendly. Is that scalable? Can we really put all data centers in France?

- **Direct Answer:** We cannot move every server to France, but we can prioritize regions with nuclear, hydro, or geothermal power.
- **Support:** The "BLOOM Lesson" isn't about France specifically; it's about grid selection. It proves that decarbonization is an infrastructure problem, not just a code problem. We should build new data centers in Quebec or Scandinavia, not coal-heavy Virginia.

15. Does your analysis account for the hardware manufacturing cost (embedded carbon)?

- **Direct Answer:** No, our analysis is strictly limited to operational carbon (Scope 2).
- **Support:** Manufacturing chips (Scope 3) adds significant carbon, but reliable data for that is even scarcer. By focusing on operational energy, we analyzed what companies can control *today* via software and scheduling.

Project Defense Questions

16. What is the single biggest weakness of your analysis?

- **Direct Answer:** The reliance on estimated benchmarks rather than raw, real-time telemetry from the GPU clusters.
- **Support:** We are analyzing the "reported" footprint, which may differ from the actual second-by-second consumption. However, this is a universal limitation in outside-in audits of closed-source technology.

17. If you had access to OpenAI's internal servers for one day, what data point would you look for first?

- **Direct Answer:** I would look for the specific "Inference-to-Training Ratio" of energy consumption.
- **Support:** Knowing exactly how much energy is spent on *users* vs. *creators* would settle the debate on where to focus regulation. Currently, we estimate inference impact, but exact internal logs would confirm if it outweighs training.

18. Why did you choose Python/Pandas for this instead of Excel?

- **Direct Answer:** Python allows for reproducible analysis and handling of logarithmic visualizations that are clumsy in Excel.
- **Support:** Using Pandas allowed us to easily filter outliers, calculate correlation coefficients programmatically, and generate complex Seaborn plots (like the regression with confidence intervals) that validate the statistics automatically.

19. How would these results change if we switched to liquid cooling?

- **Direct Answer:** The Water Intensity Ratio (slope) would drop significantly, potentially near zero for closed-loop systems.
- **Support:** Liquid immersion cooling recycles the heat transfer fluid rather than evaporating water. If the industry switched, our graph 1 (Water vs. Energy) would flatten out, decoupling water stress from energy use.

20. Summarize your entire project in one sentence. What is the takeaway?

- **Direct Answer:** We found that AI sustainability is currently a hardware and grid problem, where a few massive models drive the majority of environmental impact.
- **Support:** The data proves that software efficiency helps, but the biggest wins come from where you plug the computer in (Grid Factor) and how you cool it (Water Ratio).

5. How to Prepare for the Presentation

- **Know Your Units:** Do not mix up **kW** (speed of electricity use) with **kWh** (volume of electricity used). If you say "PaLM used 3 million kilowatts," you are wrong. It used "3 million kilowatt-hours."
- **Own the Limitations:** Don't hide the data scarcity. Start by saying, "AI transparency is low, so we worked with the best available public audits." This makes you look mature and scientific.
- **The "So What?" Factor:** Don't just point at a graph. Point at the outlier dot (PaLM) and say, "This single dot represents more energy than 100 homes use in a year. This is why we need the changes we are proposing."
- **Slide Navigation:** Know your slides cold. If a teacher asks about water, you should already be clicking to the "Water vs. Energy" slide before they finish the sentence.