## AI Model Building Process

### Parameters planning

The model needs specific, carefully chosen parameters to effectively capture the workload that an LLM (or other technologies) must manage for query processing. This workload is then expressed in terms of energy consumption, water usage, and carbon footprint.

The most important factors that directly relate to this are,

* Query length
* Query token count
* Query type (code gen, image analysis, image gen, doc analysis, etc..)
* Model name
* Region of server that responds
* No. of API calls (Gemini does this)
* Part of the day at the data center (morning/afternoon/night)
* Season present at data center area

Output of the model:

* Power consumption (in kWh)
* Water consumption (in L)
* CO2 emission (in Kg CO2)

These outputs should then be mapped to real-world examples for easy understanding.

### Data Collection for Each Parameter

To ensure a comprehensive dataset for our model, we employ a combination of simulation, publicly available data, and reasonable assumptions.

* **Query Length and Type**: Simulated using a range of predefined structures and randomly generated query formats.
* **Model Names**: Selected from a curated list of commonly used AI models.
* **Server Regions / Data Centers**: Extracted from publicly available datasets, such as those on Kaggle, or sourced from cloud provider documentation.

The challenge is in simulating the corresponding outputs for each input.

No information is available for predicting the output parameters. Cloud service providers did not provide open data on these metrics. So, we are making some assumptions and generalizations for predicting the outputs.

**DISCLAIMER:**

The predicted sustainability metrics in this model—**power consumption, water footprint, and CO₂ emissions**—are **approximations** based on simulated data and estimations. These values **do not represent real-time or provider-specific measurements** but serve as indicative benchmarks for understanding AI’s environmental impact.

### Methods used to simulate output

Using Publicly available data,

* Google Cloud carbon footprint reports
* Microsoft’s Sustainability Calculator
* Top cloud service provider’s energy and sustainability reports
* Research papers on LLM energy usage
* Stanford AI index report
* Open-source datasets available

### Research Papers to analyze

Power consumption

<https://ieeexplore.ieee.org/abstract/document/10363447>

<https://arxiv.org/abs/2205.09646>

<https://arxiv.org/abs/2407.04014>

<https://www.sciencedirect.com/science/article/pii/S2210537923000124>

<https://dl.acm.org/doi/abs/10.1145/3630106.3658542>

<https://arxiv.org/abs/2109.05472>

Carbon Footprint analysis

<https://dl.acm.org/doi/abs/10.1145/3604930.3605705>

<https://dl.acm.org/doi/fullHtml/10.1145/3603746>

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5036344>

Water consumption

<https://puiij.com/index.php/research/article/view/39>

<https://sustainability.biruni.edu.tr/sites/default/files/2024-05/Gupta%2C%20et%20al._AIs%20excessive%20water%20consumption.pdf>

<https://dl.acm.org/doi/abs/10.1145/3578337.3605121>

### Inference from research papers

For LLaMA model, which is similar to chatGPT3

* 65B (65 billion parameters, 80 layers, 8192 dimensions)
* words/sec or response/second **inversely proportional** to model parameter size
* Advanced hardware (A100 GPUs) **proportional** to energy usage but less response time
* On bare minimum hardware settings,
  + 65B: 460 W/sec (approx.)
  + 13B: 380 W/sec (approx.)
  + 7B: 200W/sec (approx.)
* On best case,
  + Approximately 1000W/sec

Model training of chatGPT - 1300MWh (approx.)

Datacenter PUE:

PUE = (FE+ITE)/ITE, FE - facility energy; ITE - information tech. Energy

PUE high during afternoons and summers

* PUE = 1.6 (afternoons) ; PUE = 1.2 (night)
* PUE = 1.3 to 1.4 during summer
* Major datacenter like google and other, PUE = 1.1

GFLOPs are more useful in compute than no. of parameters

Correlation between GFLOPs and # of parameters

* Transformers - 0.994
* CNNs - 0.772

Top 1-accuracy models have GFLOPs around 50 to 100(0.3 to 1 Joule)

Very high accuracy (85% - 90%) - around 1000 GFLOPs(30 Joules)  
  
Tokens play an important role in energy consumption

* Input tokens (8 tokens -> 80J/token ; 2^11 tokens -> 20J/token)
* Output tokens (8 tokens -> 10J/token ; 2^11 tokens -> 700J/token)

Output tokens ply crucial role in energy consumption.

### Government & Tech Reports

* **Google Cloud Sustainability Reports** (Google’s Environmental Report)
* **Microsoft Sustainability Calculator** ([Microsoft Cloud CO₂ Emissions](https://www.microsoft.com/en-us/sustainability/emissions-impact-dashboard))
* **AWS Carbon Footprint Report** (AWS Sustainability)
* **International Energy Agency (IEA) Reports** (IEA AI & Energy)