✓ Metrics

Al models process language using units called tokens, which are essentially pieces of words. To estimate the environmental impact of a model like GPT-5, the first step is to determine its energy consumption per token.

As OpenAl runs its models on Microsoft Azure data centers, this energy metric can then be combined with Microsoft's public data on water consumption and greenhouse gas emissions to calculate the model's full environmental footprint.

The University of Rhode Island's AI lab provides query resource estimates for OpenAI models GPT-5 & GPT-4o. Assuming a standard query is 1,000 tokens, we can derive the energy cost per token as follows:

• ChatGPT-5:
$$18\frac{\mathrm{Wh}}{\mathrm{query}} \times \frac{\mathrm{query}}{1000 \; \mathrm{tokens}} = \boxed{0.018 \; \mathrm{Wh/token}}$$
• GPT-4o: $0.60\frac{\mathrm{Wh}}{\mathrm{query}} \times \frac{\mathrm{query}}{1000 \; \mathrm{tokens}} = \boxed{0.0006 \; \mathrm{Wh/token}}$

The app will use GPT-5's energy estimate to calculate statistics, but will actually be running GPT-4o on the backend to conserve resources (30x more efficient)

Milliliters of Water (mL)

The Water Use Effectiveness (WUE) is an industry metric measuring a facilities rate of water consumed (L) relative to energy consumed (kWh).

Microsoft Azure WUE:
$$0.30\frac{L}{kWh} imes \frac{kWh}{1000mL} imes \frac{kWh}{1000Wh} = \boxed{0.30~mL/Wh}$$

• Water per Token:
$$0.30 rac{
m mL}{
m Wh} imes 0.018 rac{
m Wh}{
m token} = \boxed{0.054 \
m mL/token}$$

SEmissions (g CO₂e)

The Carbon Intensity Factor (CIF) is an industry metric measuring a facilities greenhouse gas emissions (kg CO₂e) relative to energy consumed (kWh).

Microsoft Azure CIF:
$$0.3528\frac{\mathrm{kg~CO_2e}}{\mathrm{kWh}} imes \frac{1000\mathrm{g}}{\mathrm{kg}} imes \frac{\mathrm{kWh}}{1000\mathrm{Wh}} = \boxed{0.3528~\mathrm{g~CO_2e/Wh}}$$

• Emissions per Token:
$$0.3528 \frac{\text{g CO}_2\text{e}}{\text{Wh}} \times 0.018 \frac{\text{Wh}}{\text{token}} = \boxed{0.0064 \text{ g CO}_2\text{e}/\text{token}}$$

USD Cost per Token (\$)

The API pricing by model are available from OpenAI:

1. **GPT-5**:

• Input:
$$\frac{\$1.250}{1 ext{M tokens}} = \boxed{\$0.00000125/ ext{token}}$$

• Cached input:
$$\frac{\$0.125}{1 \text{M tokens}} = \boxed{\$0.000000125/\text{token}}$$

• Output:
$$\frac{\$10.000}{1 \text{M tokens}} = \boxed{\$0.00001/\text{token}}$$

2. **GPT-4o**:

• Input:
$$\frac{\$2.50}{1 \text{ M tokens}} = \boxed{\$0.0000025/\text{token}}$$

• Cached input:
$$\frac{\$1.250}{1 \text{M tokens}} = \frac{\$0.00000125}{\text{token}}$$

• Output:
$$\frac{\$10.000}{1 \text{M tokens}} = \boxed{\$0.00001/\text{token}}$$

Metric	GPT-5	GPT-4o
∳ Energy	0.018 Wh	0.0006 Wh
△ Water	0.0054 mL	0.00018 mL
Emissions	0.0064 g CO₂e	0.00021 g CO₂e

© S Cost	GPT-5	GPT-4o
Input	\$0.0000125	\$0.0000125
Output	\$0.00001	\$0.00001

© Cost	GPT-5	GPT-4o
Cached	\$0.00000125	\$0.0000125

Sources:

- 1. How Hungry is Al? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference
- 2. Model Consumption & Comparison Dashboard (Very insightful!)
- 3. OpenAl will not disclose GPT-5's energy use. It could be higher than past models
- 4. Measuring datacenter energy and water use to improve Microsoft Cloud sustainability
- 5. How much carbon dioxide is produced per kilowatthour of U.S. electricity generation?
- 6. Microsoft 2024 Environmental Sustainability Report
- 7. OpenAl API Pricing