The Energy Footprint of Generative AI

Lena Pang
MATH-421 Mathematical Modeling
with Dr. Nathan Cahill
Final Project

December 9, 2024

Abstract

This paper examines the tangible and identifiable energy costs that come with the rise of generative artificial intelligence. We define a set of energy use and demand assumptions. Using data on energy expenditure from three main factors (training models, generating queries, and public demand), we created a model that predicts energy use in the future. The model estimates that by 2040, generative AI will consume a whopping 22,500 MWh of energy daily. We conclude with a discussion of possible error sources and the foreboding ethical implications of our results.

Introduction

Generative AI has taken the digital world by storm—popping up everywhere, from questionable Google search results to prompt-based music and image generators. It is nearly unavoidable. Inescapable. But massive use comes with a massive cost. Here, we determine an estimate for just *how* massive is massive.

Assumptions

For this model, we restrict the total energy consumption to that of training and queries. As such, we can define our assumptions.

(a) Training an AI model uses a finite amount of energy, and each sequential generation uses more energy than the previous.

(b) Individual queries each use a finite amount of energy.

- Text generation uses a finite amount of energy per query.
- Image generation uses a finite amount of energy per query (that is greater than the energy required to generate text).

(c) Demand for/usage of AI will continue to increase over time, but the increase in demand will slow.

Parameters

From these assumptions, we can develop a set of parameters for our governing system.

(a) Training parameters:

α_1 — energy increase factor between models

α_2 — training time factor between models

(b) Query parameters:

β — average query energy cost (wh)

$\beta_1 < 1$ — proportion of text generations

$\beta_2 < 1$ — proportion of image generations

k_{β_1} — energy usage per text (wh)

k_{β_2} — energy usage per image (wh)

where $k_{\beta_1} > k_{\beta_2}$

(d) Demand parameters:

$q(t)$ — # of daily queries as a function of time

Methodology & Realistic Parameterization

There is very little publicized research into how much energy generative AI uses. Most of what we, as the public, know is about how generative AI energy usage compares to that of algorithm-based programs. We will use this scant data to create estimates for our parameters.

(a) Training

We examine the energy consumption differences between each version of ChatGPT to estimate the energy consumption of upcoming generations.

Information on ChatGPT-2 training is not available, but ChatGPT-3 does have some publicized statistics. The International Energy Agency claims that "training a large language model like OpenAI's GPT-3, for example, uses nearly 1,300 megawatt-hours (MWh) of electricity, the annual consumption of about 130 US homes" (Calvert, 2024). Other sources calculate a similar 1,248 Mwh, based on the known duration of training and GPU specifications, so we can consider this a reliable number (TRG Datacenters). It took 34 days to train.

ChatGPT 4, however, is calculated to have taken 7,200 MWh and around 95 days to train, almost 6 times the previous generation (TRG Datacenters). It offers, supposedly, more "creative" and "problem-solving" functions, in addition to over 50 languages, image processing, and voice recognition.

GPT-2 was fully released on November 5, 2019. GPT-3 was released on November 30, 2022, almost exactly two years later. GPT-4 was released on March 14, 2023, exclusive to paying users.

We assume there is exponential growth between generations, with subsequent generations coming out around 1.5 years apart.

Thus we can define correspondingly scaled unit functions (MWh over time) for the nth generation:

$$E_3(t) = \frac{1300}{34}\left[u(t-t_3) - u(t-t_3-34)\right]$$
$$E_4(t) = \frac{7200}{95}\left[u(t-t_4) - u(t-t_4-95)\right]$$
$$E_n(t) = \frac{k_n}{c_n}\left[u(t-t_n) - u(t-t_n-c_n)\right]$$

where k_n is the amount of energy (dimensionless scalar) it took to train the model, t_n is the start date, c_n is the time it took (in days). These satisfy our known statistics:

$$\textstyle\int_{-\infty}^{\infty} E_3(t)dt = 1300 \text{ MWh}$$
$$\textstyle\int_{-\infty}^{\infty} E_4(t)dt = 7200 \text{ MWh}$$

We know the energy increases by a factor of $\alpha_1 = 6$ and the time by a factor of $\alpha_2 = 3$ between generations. So we create an approximate recursive sequence for future generations:

$$E_{n+1}(t) = 2k_n\left[u(t-t_{n+1}) - u(t-t_{n+1}-3c_n)\right]$$

(b) Queries

We first look at text data. According to the International Energy Agency, a ChatGPT 3.0 text generative request uses 2.9 watt-hours of energy—which is almost ten times the amount of energy it takes to process a Google search (Calvert, 2024). That gives us an estimate $\beta_1 = 3$ Wh. This parameter is subject to change over time, decreasing as models become more efficient or increasing as models become more complex.

Images, on the other hand, are more energy-intensive to generate. One research study tested 10 language models on NVIDIA GPUs, with the result of $2,907$ Wh on average (Jernite). So we can approximate $\beta_2 = 2907$.

We make an assumptions that the proportion of text generation to image generation is around $1000 : 1$. This number is an estimate given free ChatGPT users can only generate 2 images per day, while paid members can generate more. This gives us $k_{\beta_1} = 0.999$ and $k_{\beta_2} = 0.001$.

Finally, this gives us the expected value of energy consumption in Wh per query:

$$\begin{aligned}\beta &= k_{\beta_1}(\beta_1) + k_{\beta_2}(\beta_2)\\ &= 0.999(2) + 0.001(2907)\\ &= 5.904 \text{ Wh}\end{aligned}$$

(c) Demand

To create a predictive model of demand (a function $q(t)$ of queries per day), we find an accurate line of best fit and scale it to known data.

We determined a line of best fit by analyzing search term popularity on a scale of 0-100, from

October 2022 (the month before ChatGPT-3's release) to November 2024. The tracked terms are:

A. AI

B. ChatGPT (text generation platform)

C. Sora AI (video generation platform by OpenAI, not launched yet)

D. Adobe Firefly (image generation platform, launched in March of 2023)

E. Generative AI

F. OpenAI

The scatterplots on the next page (Figure 1) display the datasets with lines of best fit that are: 1. linear, 2. logarithmic, and 3. exponential.

We also added a line of the combined data normalized to the same 0-100 scale. To make the combined value more accurate, we doubled the weights of AI, ChatGPT, and Generative AI. This prevented skewing from the other less significant factors. Specifically:

$$\text{Combined}_i = 2A_i + 2B_i + C_i + D_i + 2E_i + F_i$$

While all results are fairly similar, the logarithmic line of best fit consistently has the greatest R^2 value for both the combined and individual datasets. This aligns with the idea that interest will continue to increase over time, but the rate at which new people become interested declines.

Now we have to scale our equation; in other words, find the parameters of our logarithmic equation so the numbers match reality. For this, we utilize search and page analytics.

Five days after launch, ChatGPT hit 1,000,000 visits. In November 2024 alone, ChatGPT was visited several billion times, though the true number is unreliable: Semrush says 4.8 billion, SimilarWeb says 3.8 (Semrush, SimilarWeb).

1,000,000 visits in the first five days to around 4 billion visits in the most recent 30 days–that gives us two points: 200,000 per day at Day "1", to 140,000,000 per day two years later.

So we can make a function q_0 of visits (in millions, for simplicity) per day, fitting a logarithmic line to our known restrictions:

$$\int_{700}^{730} q_0(t)dt \approx 4000$$

$$q_0(1) \approx 0.2$$

$$q_0(730) \approx 140$$

$$q_0(t) = 52\ln(0.2t + 10) - 120$$

Now we scale this up by the number of queries per visit to produce $q_(t)$. The exact number of queries per visit is somewhat of a trade secret, which leaves us to speculation. We make an assumption that each visit averages 15 requests, accounting for the many users who make only a few queries, and for the few who make many.

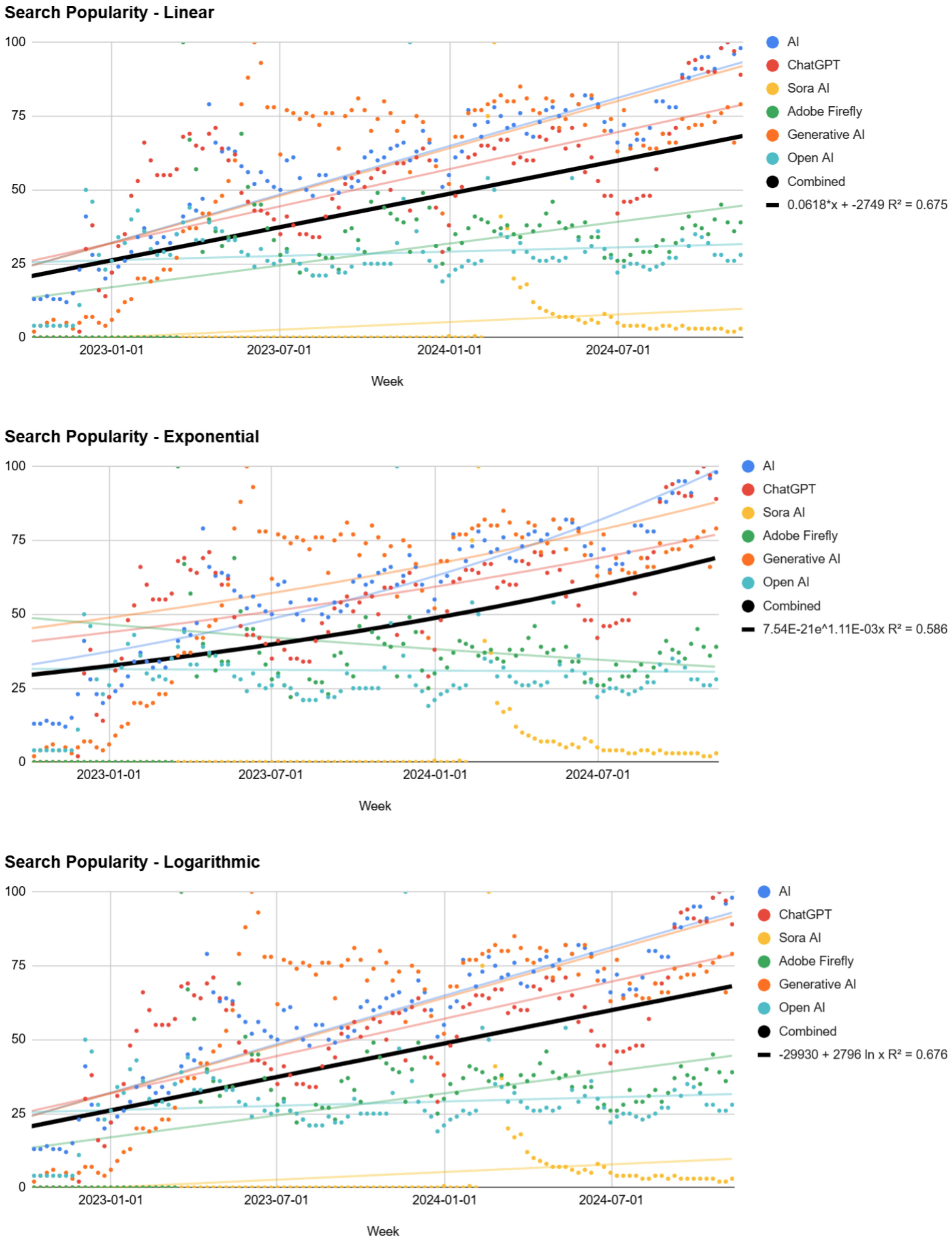
So we get $q(t)$, the number of queries (in millions) as a function of time:

$$q(t) = 15q_0(t)$$

$$= 15\left[52\ln(0.2t + 10) - 120\right]$$

$$= 780\ln(0.2t + 10) - 1800$$

Figure 1: Determining a Line of Best Fit



Search Popularity - Linear

Search Popularity - Exponential

Search Popularity - Logarithmic

Data from Google Trends (https://www.google.com/trends).

Governing System & Model

Having established realistic parameters and determined our sub-equations, we can establish an overarching system of energy consumption.

The total energy is the number of queries times the average Wh per query, added to the individual training unit functions we defined.

Thus we get:

$$E(t) = \beta q(t) + \sum_{n=3}^{\infty} E_n(t)$$

$$= 10^6 \times 5.904\,[780 \ln(0.2t + 10) - 1800] + \mathrm{E}_2(t) + E_3(t) + \dots$$

Then the cumulative energy usage, which can evaluate and plot numerically, is:

$$\int_0^t \Sigma E(t)dt = \int_0^t q(t) + \sum_{n=3}^{t_n>t} E_n(t)dt$$

We plot the predictive model and the cumulative model (starting November 5, 2022) below, with a hypothetical $E_5(t)$ in early 2025. Evidently, the overall movement aligns with our original assumptions.

Solution

The model records and predicts the energy consumption (in MWh) per day and cumulatively:

year	per day	cumulative
2023	3,600	0.05×10^7
2024	10,400	0.3×10^7
2025	12,500	0.71×10^7
2026	14,600	1.23×10^7
⋮	⋮	⋮
2030	18,000	4×10^7
⋮	⋮	⋮
2035	21,000	7×10^7
⋮	⋮	⋮
2040	22,500	11×10^7

That's around $1,368,750$ MWh in 2024 alone—the same amount of energy that it takes to power 120 thousand houses for an entire year.

The cumulative number is even worse. 11×10^7 MWh, or $110,000,000,000,000$ (yes, that's 110 trillion) Wh. That's about one million houses powered for a year, or about the amount of energy that Argentina or Sweden uses in a year.

Figure 2: Predictive and Cumulative Models (Long-Term)

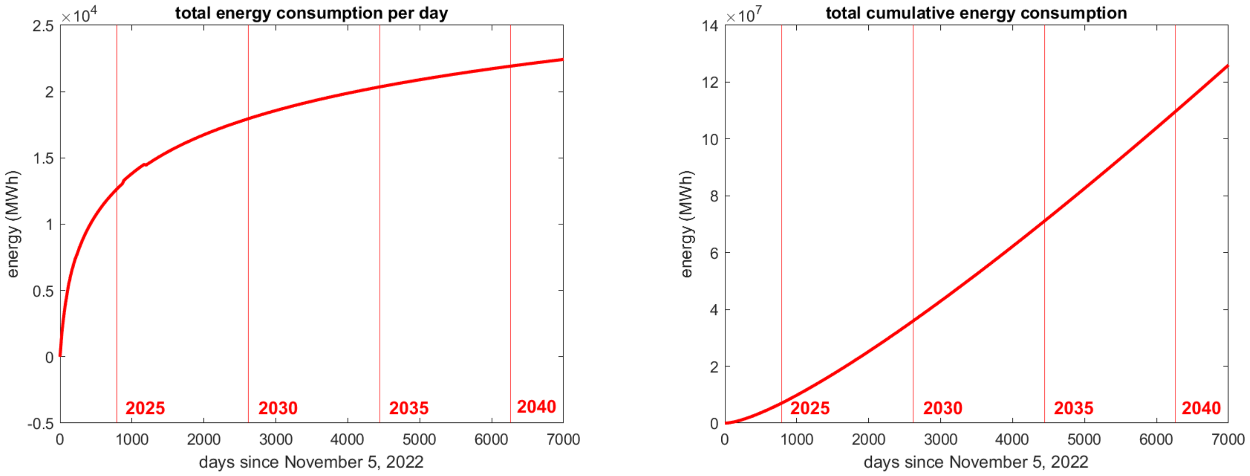
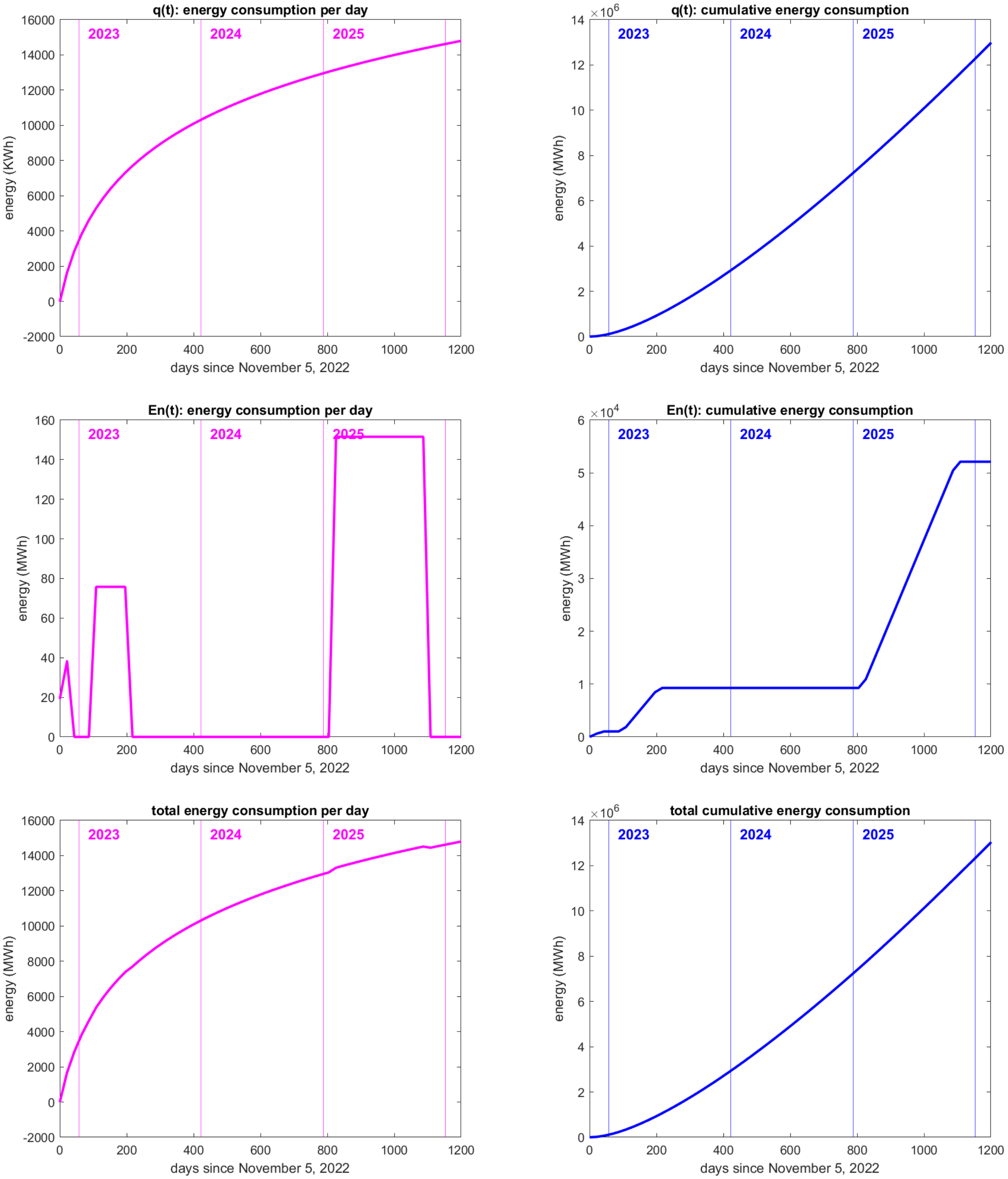


Figure 3: Predictive and Cumulative Models (Short-Term)



Error

We used several assumptions to narrow our model down to four identifiable and quantifiable factors. Specifcally, we only used publicly available estimates, and we did not calculate active data storage/retrieval (because there are no concrete numbers available).

Unfortunately, that likely means this model undershoots the real energy consumption amount, anywhere from a "negligible" difference to possibly a drastic extent. It is currently not feasible to approximate the extent to which AI companies cover their tracks for the sake of profit.

Ethics of Energy

There is AI that is worth the cost: scientific anomaly-detecting, predictive, integrated models; AI that can detect cancer; AI that can do what humans can't.

Generative AI is not that. It is a scapegoat, a replacement, for human creativity, thought, function, and connection. There is no justification for the immense energy costs. (And that's not even mentioning the non-renewable yet reliable energy sources used, nor the plethora of copyright infringements, nor the outsourced slave labor for filtering models, nor the water contamination from data center cooling.)

Nevertheless, AI has a chokehold on the future of technology. It is direly important to regulate and restrict usage for the sake of our critical and creative minds, and for the sake of our one-and-only planet Earth.

Conclusion

We analyzed energy usage and trend data to create a predictive model for the energy footprint of generative AI. We predict that the daily energy consumption will be around 18,000 MWh by 2030, and around 22,500 MWh by 2040—likely underestimates due to lack of data availability.

We must decide what price we are willing to pay for generative AI access. What use is a stochastic pseudo-intelligent machine if it kills us all?

References

Calvert, B. (2024). AI already uses as much energy as a small country. It's only the beginning. *Vox*. https://www.vox.com/climate/2024/3/28/24111721/climate-ai-tech-energy-demand-rising

Jernite, Y., Luccioni, A. S., Strubell, E. Power Hungry Processing: Watts Driving the Cost of AI Deployment? https://arxiv.org/pdf/2311.16863

SimilarWeb. (2024). chatgpt.com Traffic & Engagement Analysis. https://www.similarweb.com/website/chatgpt.com/#ranking

Semrush. (2024). Traffic Analytics: chatgpt.com. https://www.semrush.com/analytics/traffic/overview/?q=chatgpt.com&searchType=domain

TRG Datacenters. AI Chatbots: Energy usage of 2023's most popular chatbots (so far). https://www.trgdatacenters.com/resource/ai-chatbots-energy-usage-of-2023s-most-popular-chatbots-so-far/

Appendix (Figures 2 & 3)

```
%% SETUP ----------------------------------------------------------------------------

% domain
x = linspace(0,6500,300);

% queries
qt = 1000000*5.904*(780*log(0.2*x+10)-1800)/1000000;

% training
t3 = 0;
t4 = 104;
t5 = 822;
e4 = 7200/95*(heaviside(x - t4) - heaviside(x - t4 - 95));
e3 = 1300/34*(heaviside(x - t3) - heaviside(x - t3 - 34));
e5 = 2*7200/95*(heaviside(x - t5) - heaviside(x - t5 - 95*3));
en = e3 + e4 + e5;

% everything
et = qt + e3 + e4 + e5;

%% SHORT TERM -----------------------------------------------------------------------

% q(t)
figure;
plot(x,qt,’LineWidth’,2,’Color’,"magenta");
xline(57,"magenta"); % Jan 2023
xline(422,"magenta"); % Jan 2024
xline(788,"magenta"); % Jan 2025
xline(1153,"magenta"); % Jan 2026
xlim([0 1200]);
title(’q(t): energy consumption per day’);
xlabel(’days since November 5, 2022’);
ylabel(’energy (KWh)’);
saveas(gcf,’q(t) short.png’);

% en(t)
figure;
plot(x,en,’LineWidth’,2,’Color’,"magenta");
xline(57,"magenta"); % Jan 2023
xline(422,"magenta"); % Jan 2024
xline(788,"magenta"); % Jan 2025
xline(1153,"magenta"); % Jan 2026
xlim([0 1200]);
title(’En(t): energy consumption per day’);
xlabel(’days since November 5, 2022’);
ylabel(’energy (MWh)’);
saveas(gcf,’en(t) short.png’);

% e(t)
figure;
plot(x,et,’LineWidth’,2,’Color’,"magenta");
xline(57,"magenta"); % Jan 2023
xline(422,"magenta"); % Jan 2024
xline(788,"magenta"); % Jan 2025
xline(1153,"magenta"); % Jan 2026
xlim([0 1200]);
title(’total energy consumption per day’);
xlabel(’days since November 5, 2022’);
ylabel(’energy (MWh)’);
saveas(gcf,’e(t) short.png’);

%% SHORT TERM CUMULATIVE -------------------------------------------------------

% q(t)
figure;
ct = cumtrapz(x,qt);
plot(x,ct,’LineWidth’,2,’Color’,"blue");
xline(57,"blue"); % Jan 2023
xline(422,"blue"); % Jan 2024
xline(788,"blue"); % Jan 2025
xline(1153,"blue"); % Jan 2026
xlim([0 1200]);
title(’q(t): cumulative energy consumption’);
xlabel(’days since November 5, 2022’);
ylabel(’energy (MWh)’);
saveas(gcf,’q(t) cumulative.png’);

% en(t)
figure;
ct = cumtrapz(x,en);
plot(x,ct,’LineWidth’,2,’Color’,"blue");
xline(57,"blue"); % Jan 2023
xline(422,"blue"); % Jan 2024
xline(788,"blue"); % Jan 2025
xline(1153,"blue"); % Jan 2026
xlim([0 1200]);
title('En(t): cumulative energy consumption');
xlabel('days since November 5, 2022');
ylabel('energy (MWh)');
saveas(gcf,'en(t) cumulative.png');

% e(t)
figure;
ct = cumtrapz(x,et);
plot(x,ct,'LineWidth',2,'Color',"blue");
xline(57,"blue"); % Jan 2023
xline(422,"blue"); % Jan 2024
xline(788,"blue"); % Jan 2025
xline(1153,"blue"); % Jan 2026
xlim([0 1200]);
title('total cumulative energy consumption');
xlabel('days since November 5, 2022');
ylabel('energy (MWh)');
saveas(gcf,'e(t) cumulative.png');

%% LONG TERM -----------------------------------------------------------------------

% update x values for wider domain
x = linspace(0,7000,300);

% daily
figure;
plot(x,et,'LineWidth',2,'Color',"red");
xline(788,"red"); % Jan 2025
xline(2614,"red"); % Jan 2030
xline(4440,"red"); % Jan 2035
xline(6266,"red"); % Jan 2040
xlim([0 7000]);
title('total energy consumption per day');
xlabel('days since November 5, 2022');
ylabel('energy (MWh)');
saveas(gcf,'daily long term.png');

% cumulative
figure;
ct = cumtrapz(x,et);
plot(x,ct,'LineWidth',2,'Color',"red");
```

```
xline(788,"red"); % Jan 2025
xline(2614,"red"); % Jan 2030
xline(4440,"red"); % Jan 2035
xline(6266,"red"); % Jan 2040
xlim([0 7000]);
title('total cumulative energy consumption');
xlabel('days since November 5, 2022');
ylabel('energy (MWh)');
saveas(gcf,'cumulative long term.png');
```