# Economic Analysis of LLM Inference Costs (2023–2026)

## Introduction

Large Language Models (LLMs) such as those developed by **OpenAI** and **Anthropic** are experiencing massive adoption. Yet, their operation relies on an expensive infrastructure: high-end GPUs, high electricity consumption, and limited energy optimization in single-agent usage.

The objective of this project is to:

- Assess the **evolution of inference costs** over time.
- Identify **break-even thresholds** for different subscription profiles.
- Propose a **5–10 year projection** of costs and the subscription prices required for profitability.

---

## Methodology

1. **Data sources**:
    - **EIA** (U.S. Energy Information Administration) for commercial electricity prices.
    - GPU pricing estimates (H100, L4) via **public markets** and manual overrides.
    - Internal calculations: energy consumption, PUE (Power Usage Effectiveness), throughput (tokens/sec).
2. **ETL pipeline**:
    - Extraction (Python, EIA API + GPU scrapers).
    - Transformation (cleaning, temporal harmonization, GPU/electricity cost integration).
    - Loading into **PostgreSQL** via `\copy`.
3. **Analysis**:
    - Calculation of **cost per million tokens** (electricity + GPU).
    - Definition of subscription profiles:
        - **Lite**: 200k tokens/month.
        - **Standard**: 1M tokens/month.
        - **Pro**: 5M tokens/month.
    - Application of a **70% target margin** to estimate break-even prices.

---

# Analysis

## 1. Observed costs (2023–2026)

- The **average cost per million tokens** ranges between **1.8 and 2.2 USD** depending on the period.
- Costs fluctuate with:
  - Changes in electricity prices (EIA source).
  - Adjustments to GPU hourly rates (H100, L4).

## 2. Break-even thresholds

- **Lite (200k tokens/month)**: break-even ≈ **1.2–1.4 USD/month**.
- **Standard (1M tokens/month)**: break-even ≈ **6–7 USD/month**.
- **Pro (5M tokens/month)**: break-even ≈ **31–34 USD/month**.

These levels remain **far below current subscription prices** (ChatGPT Plus: 20 USD/month, Claude Pro: 20 USD/month).
This indicates that companies are massively subsidizing access, despite high costs (and here we only consider electricity—other factors also come into play).

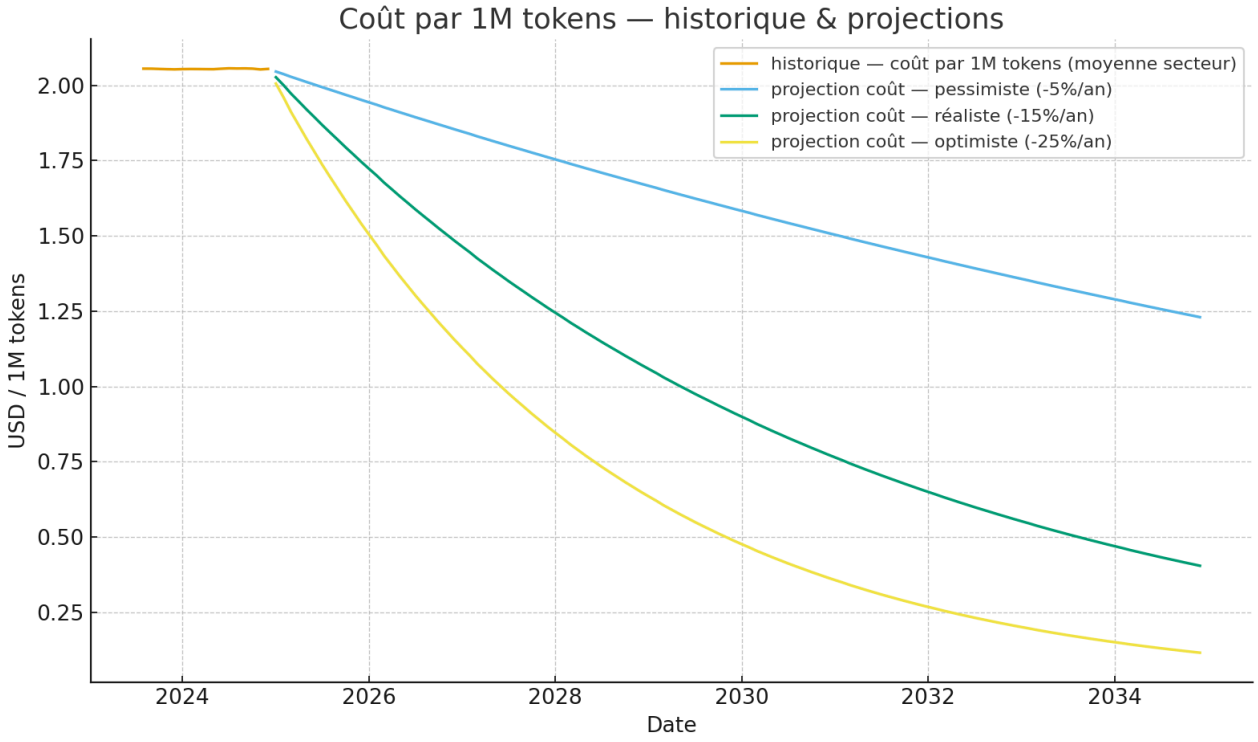
## 3. Structural deficit

- Even without including **R&D, salaries, servers, and storage**, operations remain **unprofitable**.
- Current subscription prices do not cover the real marginal cost of inference.

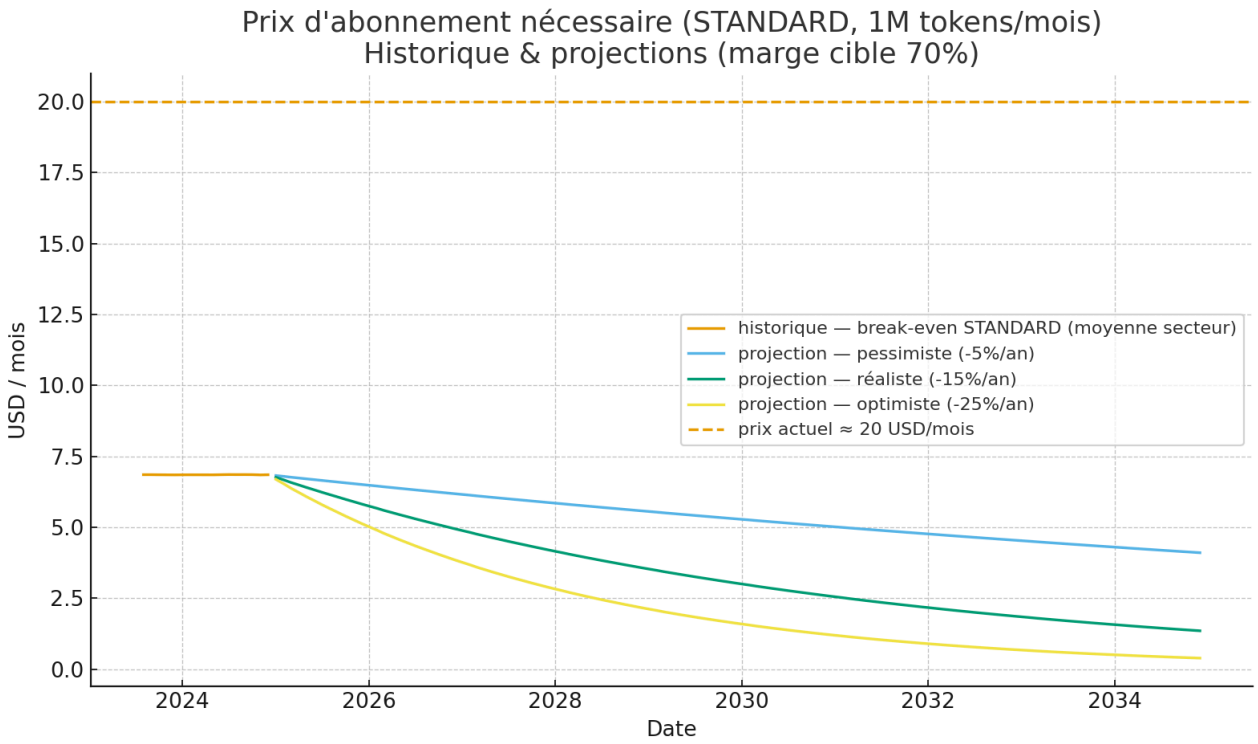
---

# Visualizations (5–10 years)

1. **Cost per million tokens curve (2023–2026)**.



2. **Hypothetical subscription price curve** required to reach profitability.




*(Charts generated via matplotlib, available in the Python notebook.)*

---

# Discussion

Why do OpenAI and Anthropic continue despite the losses?

- **Network effect**: more users = more data = better models.
- **Strategic race**: AI is a winner-takes-all sector; leadership matters more than immediate profitability.
- **Massive subsidies**: Microsoft, Google, and Amazon heavily fund these players.
- **Bet on the future**: GPU/energy costs may decrease, or subscription prices may rise.
- **Market entrenchment**: becoming indispensable.

In the long run, it is likely that:

- Subscriptions will **gradually increase** (25–40 USD/month).
- Offers will become more **segmented** (Lite, Pro, Enterprise).
- Companies will rely on **ancillary revenues** (API, integrations, SaaS products).

---

# Conclusion & outlook

- LLMs are currently operated at a loss, even when considering only **GPU + electricity**.
- Long-term economic viability will depend on:
  - **Higher subscription prices**.
  - **Energy optimization** (PUE, specialized chips).
  - **Scale effects** and throughput improvements.

- Our projections indicate a **necessary increase in subscription fees within 5–10 years**, otherwise losses will become unsustainable.

**Next steps:**

- Extend projections to 2030–2035 with optimistic/pessimistic scenarios.
- Broaden the analysis to other players (Mistral, Meta, Google DeepMind).
- Simulate the impact of a carbon tax on the final cost of LLMs.