Data Insight Mar. 12, 2025

LLM inference prices have fallen rapidly but unequally across tasks

Cite



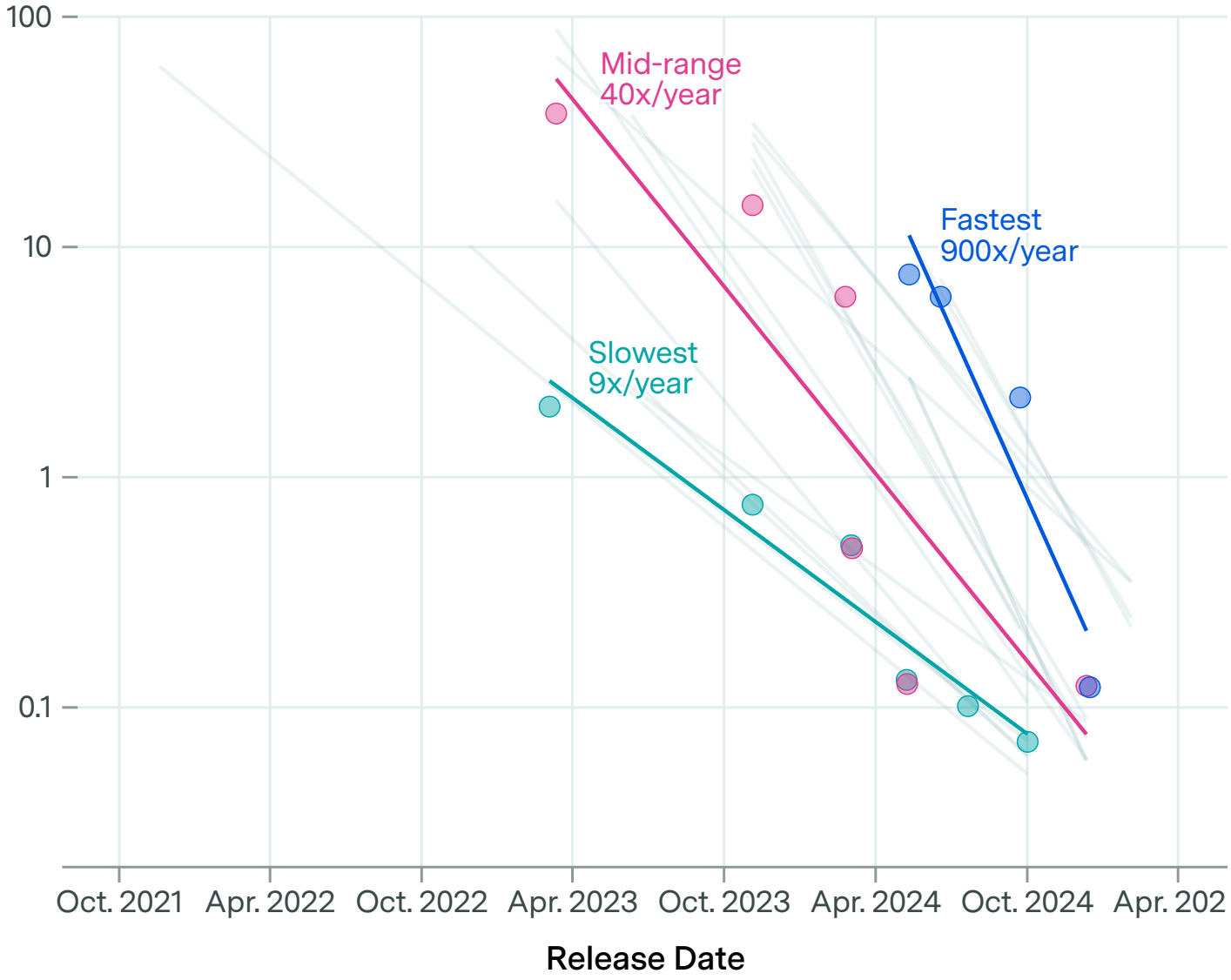
By Ben Cottier, Ben Snodin, David Owen, and Tom Adamczewski

The inference price of LLMs has fallen dramatically in recent years. We looked at the results of state-of-the-art models on six benchmarks over the past three years, and then measured how quickly the price to achieve those milestones has fallen. For instance, the price to achieve GPT-4's performance on a set of PhD-level science questions fell by 40x per year. The rate of decline varies dramatically depending on the performance milestone, ranging from 9x to 900x per year. The fastest price drops in that range have occurred in the past year, so it's less clear that those will persist.

View

All benchmarks | MMLU (General knowledge) | GPQA Diamond (Ph.D. level science questi

Price (USD per million tokens)



- GPT-3.5 Turbo level or better on general knowledge (MMLU)
- GPT-4 level or better on Ph.D. level science questions (GPQA)
- GPT-4o level or better on Ph.D. level science questions (GPQA)
- Other benchmarks and performance levels

Data source: Epoch AI, Artificial Analysis

EPOCH AI | CC-BY

There are many potential reasons for these price drops. Some are well-known, such as models becoming smaller and hardware becoming more cost-effective, while other important factors might be difficult to determine from public information.

Epoch's work is free to use, distribute, and reproduce provided the source and authors are credited under the Creative Commons BY license.

Learn more about this graph

The dataset for this insight combines data on large language model (LLM) API prices and benchmark scores from Artificial Analysis and Epoch AI. We used this dataset to identify the lowest-priced LLMs that match or exceed a given score on a benchmark. We then fit a log-linear regression model to the prices of these LLMs over time, to measure the rate of decrease in price. We applied the same method to several benchmarks (e.g. MMLU, HumanEval) and performance thresholds (e.g. GPT-3.5 level, GPT-4o level) to determine the variation across performance metrics.

Note that while the data insight provides some commentary on what factors drive these price drops, we did not explicitly model these factors. Reduced profit margins may explain some of the drops in price, but we didn't find clear evidence for this.

Code to reproduce our analysis can be found here.

Data +

Analysis +

Assumptions +

Download this data

Lowest inference prices at fixed performance
CSV, Updated Nov. 20, 2025

Inference prices of other models

CSV, Updated Nov. 20, 2025

Explore this data



Capabilities & benchmarking

Benchmark results featuring the performance of leading AI models on challenging tasks.

Related topics

Finances

Scaling