ZEFFER BUSINESS INTELLIGENCE

AI ECONOMICS / OCTOBER 26, 2026

THE PRICE OF AI IS FALLING.

Cheaper tokens. More open models. A tougher test for pricing power.



01 TOKEN PRICES KEEP DROPPING

−41%

Effective price per million tokens

Ramp's blended-price measure [1]

Mar 2026	$1.15
Sep 2026	$0.68

March peak → early September 2026.
A blended average, not the price of every model.

02 LOW-COST MODELS GAIN USERS

DeepSeek in production

~13% → >40%

JANUARY 2026 — SEPTEMBER 2026

Share of U.S. startups on OpenRouter using DeepSeek in production. [2]

A platform sample of startups, not all U.S. firms.

03 THE USAGE-REVENUE GAP

Open-weight models on Vercel

Token share	71.5%
Spend share	23%

Most tokens.
A smaller share of dollars.

September 2026 • AI Gateway [3]
Closed models retained 77% of estimated spending.

04 CHEAP, OPEN AND FREE ARE DIFFERENT

Budget APIs

Pay per token.
Lower-cost models can handle many routine workloads. [1,2]

Open weights

Some weights are free to use.
Self-hosting still needs hardware and operations.

Free tiers

No usage charge within limits.
Availability and rate limits vary by provider. [4]

ZEFFER VIEW

More tokens do not automatically mean more profit.

50% lower price × 2× volume = unchanged revenue

Illustrative math; assumes the same mix of tokens. Profit also depends on delivery costs.

Demand can win: Vercel gateway spending still rose ~13× from Dec 2025 to Sep 2026. [3]

SOURCES

[1] Ramp AI Index, Sep 9, 2026 • ramp.com/data/ai-index-sept-2026
[2] OpenRouter, The Token State of Play, Sep 22, 2026 • openrouter.ai/data/newsletter
[3] Vercel AI Gateway Production Index, Oct 15, 2026 • vercel.com/blog
[4] OpenRouter Free Variant documentation, accessed Oct 26, 2026 • openrouter.ai/docs

Platform samples, not global market shares. Vercel spending is estimated at list prices. Token mix and caching affect averages.

© 2026 Zeffer Business Intelligence • Data. Context. Perspective.

zeffer.ai