Why Stochastic Reasoning Makes Revshare the Only Economically Compatible Model for LLMs

A Data-Driven Revenue Hypothesis for OpenAI

By Nuno Lopes Bsc.

Executive summary

The internet's dominant business model has, for two decades, been built on advertising, attention and hope. Advertisers buy clicks in the hope that some proportion will convert. Platforms optimise auctions and engagement. Users pay indirectly, with time and data. This structure created trillion-dollar companies → but the model is now showing structural strain: rising customer acquisition costs (CAC), increasing auction complexity, and diminishing returns for smaller businesses.

Large language models (LLMs) such as OpenAI's models are entirely different economic objects. They engage in stochastic reasoning over long context windows, gather rich detail about users' intentions, and can execute complex, multi-step tasks. This architecture is fundamentally mis-aligned with ad auctions, which reward bidding power rather than reasoning quality.

This paper develops a data-driven hypothesis:

Outcome-based revenue sharing (revshare) is the only monetisation model that is economically compatible with frontier LLMs at scale.

Using recent estimates of OpenAI's revenue and compute costs, alongside benchmarks from Google Ads and global ad-spend forecasts, the paper argues that:

OpenAI has reached an annualised revenue run-rate of around US$10–13 billion by mid-2025, up from ~US$2–4 billion in 2023–24 (Reuters, 2025; Entrepreneur, 2025).

Its compute costs are extraordinary: leaked documents and independent analysis suggest cumulative spend of over US$12 billion on Azure inference alone since 2024, with total 2024 compute (training + inference + research) around US$5.8 billion and 2025 inference spend through Q3 already exceeding US$8.6 billion (Financial Times, 2025; The Register, 2025; Epoch, 2025b).

Average Google Ads cost-per-click (CPC) in 2025 is around US$5.26, with average cost-per-lead (CPL) ~US$70, and multi-year upward pressure on CPC and CPL for many verticals (WordStream, 2025b; LocaliQ, 2025).

Global advertising spend is projected to exceed US$1 trillion in 2025, with digital now the majority of spend and on track to dominate by 2030 (WARC, 2025; eMarketer, 2025).

Against this backdrop, the paper models revshare economics at the level of a single commercial conversation and shows that:

A high-intent AI-mediated session can realistically generate US$3–4 of gross profit per successful conversation under plausible assumptions about order value, revshare percentage, and conversion rate.

At relatively modest volumes → say 100 million high-intent sessions per month → this implies US$4–5 billion per year in gross profit, enough to materially offset current inference burn.

Unlike pay-per-click (PPC), revshare pays the model only when the user's problem is actually solved → which reinforces rather than corrupts the model's stochastic reasoning.

The conclusion is that LLMs monetise understanding rather than attention. Revshare is the only model that scales with that property without undermining reasoning quality, user trust or the economics of smaller businesses.

1. Introduction: from attention to outcomes

The web's economic history since the early 2000s can be described, in simple terms, as the rise of the attention → ad model. Platforms such as Google and Meta built systems that:

Captured user attention (search, feeds, video).

Sold that attention to advertisers via auctions.

Optimised for clicks and impressions as proxies for value.

The result was an ecosystem where visibility is purchased. The more a merchant or brand is willing to bid, the more often they are shown. Economically, this model treats attention as the scarce resource.

Several cracks in this model are now apparent:

Rising CAC: in many sectors, digital CAC has increased faster than revenue growth, squeezing margins and forcing brands to rely on constant fundraising.

Auction complexity and opacity: smaller businesses struggle to compete effectively in PPC auctions and lack the analytical capability and capital to run continuous bid-optimisation.

User fatigue: users increasingly distrust ad-heavy environments and look for more direct, less noisy ways to find what they want.

Enter large language models. Rather than presenting a ranked list of links, an LLM can:

Interpret long, messy, emotionally coloured queries.

Ask clarifying questions.

Evaluate trade-offs.

Call tools and APIs.

Execute transactions.

The user experience shifts from "search, click, hope" to "ask, reason, solve". That change is not cosmetic; it implies an entirely different revenue logic. An LLM that is artificially biased by ad auctions will produce worse reasoning, lose trust, and ultimately convert less.

This leads to the central question of the paper:

What kind of business model can fund multi-billion-dollar compute costs without corrupting the stochastic reasoning that makes LLMs valuable?

2. The cracks in Google’s economic foundation

2.1. Rising costs and fragile margins

Benchmarks from WordStream and other PPC analytics providers indicate that Google Ads costs have risen materially over recent years. For 2025, the average CPC across all industries is estimated at around US$5.26, with average CPL about US$70.11 (WordStream, 2025b).

Other analyses place search CPC in the US$2.69–5.26 range depending on sector, with display CPC around US$0.63, again with a general upward drift over time (StoreGrowers, 2025).

At the same time, global ad spend is on track to exceed US$1.0–1.1 trillion in 2025, with online media accounting for the majority of that growth (WARC, 2025; eMarketer, 2025).

For advertisers, the unit economics of search increasingly look like this:

Higher CPC and CPL

Lower marginal returns on incremental spend

Increasing dependence on sophisticated bidding, creative testing, and data science

Well-capitalised incumbents can absorb these costs and build internal teams to manage them. Many small and mid-sized businesses cannot.

2.2. Capital as a precondition for visibility

The ad auction system embeds a capital requirement into discovery. To even test the waters in a competitive vertical, an SME might need a monthly ad budget in the US$1,000–10,000 range and accept a high degree of volatility in performance (WordStream, 2025b).

This means:

New entrants must fund experimentation before they can know whether a campaign will work.

Incumbents with strong cash flows can outbid competitors to defend category visibility.

Visibility is not purely a function of product quality or relevance; it is heavily mediated by ability to pay.

The result is an ecosystem where Google's moat is not simply "better search"; it is the fact that everyone must pass through a paid gate to be seen at scale. From a market-design perspective, this is a classic case of allocation by auction rather than merit.

2.3. Search intent erosion and behavioural shifts

There are also early signs of structural change in user behaviour:

Younger users increasingly use TikTok, Reddit, or other social platforms as discovery tools for products, travel and tutorials.

“Zero-click” search experiences → where Google answers directly in the SERP → reduce clicks to third-party sites and erode the value of organic SEO.

LLM chat interfaces are starting to be used as first-line query tools for coding, research, summarisation, and increasingly commerce and local services.

While Google remains dominant in raw query volume, the marginal value of an additional search query may be declining as users fragment across multiple channels.

3. OpenAI’s economics: a very expensive brain

3.1. Revenue growth at unprecedented speed

Independent analyses suggest OpenAI's revenue growth has been extremely rapid. Epoch estimates that annualised revenue grew from roughly US$2 billion at the end of 2023 to US$13 billion by August 2025, implying a compound growth rate of more than 3× per year (Epoch, 2025a).

Reuters reported that by June 2025, OpenAI's revenue run-rate had reached US$10 billion, up from US$5.5 billion in December 2024 (Reuters, 2025). Entrepreneur further notes that revenue in the first half of 2025 exceeded the whole of 2024, with internal targets for 2025 around US$13 billion (Entrepreneur, 2025).

In pure revenue-growth terms, OpenAI is one of the fastest-scaling software businesses in history.

3.2. Compute costs: the other side of the equation

The cost side, however, is enormous. Multiple sources indicate that:

OpenAI spent around US$3 billion on training compute, US$1.8 billion on inference, and US$1 billion on research compute in 2024 (Epoch, 2025b).

Leaked documents suggest that by Q3 2025, inference costs alone had reached roughly US$8.67 billion year-to-date, up from US$3.76 billion in 2024 (Financial Times, 2025).

The Financial Times reports that OpenAI appears to have spent more than US$12.4 billion on Azure inference compute over the last seven calendar quarters (Financial Times, 2025).

This implies:

Very high fixed and quasi-fixed costs to train and operate frontier models.

A steadily rising variable cost per token or per request as usage grows, even as per-unit costs fall with optimisation.

A business that, in 2024–25, is likely not only unprofitable but heavily reliant on continued capital inflows and strategic cloud partnerships.

In short: OpenAI has built a global reasoning substrate that is both economically powerful and financially hungry. Any sustainable business model must be capable of:

Monetising a significant fraction of high-value usage.

Doing so without undermining the quality of that reasoning.

4. Why stochastic reasoning and ad auctions are incompatible

4.1. How LLMs actually reason

At a high level, an LLM works by:

Taking a long context window of tokens (user messages, prior conversation, tool outputs).

Passing them through many layers of a neural network trained to predict the next token.

Sampling from the resulting probability distribution in a stochastic way, potentially guided by decoding strategies (temperature, nucleus sampling, etc.).

In a commerce or decision-support context, this process:

Evaluates multiple candidate options implicitly in vector space.

Integrates hard constraints (budget, dimensions, delivery time) with soft preferences (style, brand trust, risk tolerance).

Uses multi-turn interaction to refine its understanding of what the user actually wants.

The result is a high-resolution embedding of user intent → far richer than a two- or three-word search query.

4.2. What happens if you inject bidding into this process?

If one were to graft a traditional ad auction onto this reasoning process → for example, by letting merchants bid to appear in the model's outputs → the following problems arise:

Bias in candidate selection: the model is nudged towards products from high bidders even when they are not the best fit.

Erosion of trust: users rapidly learn that recommendations are "sponsored", which undermines the conversational contract.

Distortion of training signals: if outputs are influenced by bids, feedback data no longer reflects pure user satisfaction but a mix of satisfaction and auction outcomes.

Lower long-term conversion: worse alignment between recommendations and user needs leads to more returns, cancellations, and dissatisfaction.

In other words, ad auctions contaminate the very property that makes LLMs valuable: unbiased, context-sensitive reasoning.

By contrast, a model that is paid only when the user's problem is actually solved → i.e. when a transaction completes, a booking is confirmed, or a service is delivered → has every incentive to optimise its chain of thought for correctness and fit, not for bidder revenue.

5. A quantitative model of revenue per conversation

To evaluate the viability of revshare as a primary revenue model, consider a simple but realistic model of a single commercial conversation.

5.1. Define the scenario

Assume:

A user engages the agent with a high-intent query, e.g. “Find me a mid-range pair of running shoes, under £120, for overpronation, UK size 8, delivered this week.”

The model interacts for several turns, clarifies constraints, and calls marketplace APIs (e.g. Shopify, Etsy, a first-party store).

It returns a shortlist of options, explains trade-offs, and the user chooses one.

The transaction is processed via a payments partner (e.g. Stripe) and the agent platform receives a revshare percentage τ of the order value.

Let:

AOV = average order value (in USD).

τ = revshare rate (fraction of order value paid to the platform).

p = probability that a high-intent session leads to a transaction (conversion rate).

C_inf = marginal inference cost for that session (compute + overhead).

5.2. Parameter assumptions

Become a member

We choose conservative, order-of-magnitude assumptions:

AOV = US$90 (mid-range apparel, electronics accessories, or similar).

τ = 12% (in line with many affiliate / marketplace commission structures).

p = 35% for genuinely high-intent sessions where the user has come to “buy with guidance”, not just browse. This is high relative to conventional e-commerce, but plausible given funnel compression and strong intent.

C_inf = US$0.05 per multi-turn session, assuming inference and orchestration costs decline over time with model optimisation and hardware improvements.

5.3. Expected revenue and margin

Then:

E[Revenue per session] = $p \times AOV \times \tau$

E[Gross profit per session] = $p \times AOV \times \tau - C_{inf}$

Plugging in values:

Revenue $\approx 0.35 \times 90 \times 0.12$ = US$3.78

Gross profit $\approx 3.78 - 0.05$ = US$3.73

Even if we halve the conversion rate (p = 0.175) or reduce AOV, we still obtain US$1.5–2+ of margin per high-intent session.

5.4. Scaling to platform level

Suppose that, out of the hundreds of millions of weekly active users OpenAI reportedly has, only a small fraction of sessions are of this high-intent commercial type → say 100 million such sessions per month.

Approximate annual gross profit:

100m × 3.73 × 12 ≈ 4.48 billion USD per year

At 200 million high-intent sessions per month, this rises to ~US$9 billion per year.

Given current estimates of OpenAI's annualised revenue (US$10–13 billion in 2025) (Reuters, 2025; Entrepreneur, 2025) and inference costs (~US$3.8 billion in 2024, already >US$8.6 billion by Q3 2025) (Epoch, 2025b; Financial Times, 2025), this simple model shows that a modest penetration of agent-mediated commerce could:

materially offset current inference burn;

eventually transform the cost structure into a high-margin flywheel if revenue scales faster than cost per session.

Crucially, this revenue is directly tied to model performance: if the agent recommends poor products, conversion falls and revshare collapses.

6. Fairness and incentive design

6.1. Revshare vs PPC as allocation mechanisms

From a market-design perspective, PPC and revshare embody different allocation principles:

PPC auctions allocate visibility based on willingness to pay per click.

Revshare allocates income based on ability to satisfy user intent, subject to the platform’s ranking function.

In revshare:

There is no upfront cost for the merchant; they pay only on success.

The platform optimises for expected value of outcome, not expected value of click.

Small merchants can, in principle, rank highly if their products best match user needs, even with limited capital.

This has clear fairness and efficiency advantages for fragmented markets with long-tail sellers.

6.2. Why small businesses are structurally advantaged under revshare

Consider a niche seller on a marketplace like Etsy or a small Shopify store:

Under PPC, they must bid against global brands for keywords, often at CPCs of US$3–6 or more (WordStream, 2025b; StoreGrowers, 2025).

Under revshare with an LLM agent, they simply expose structured product data (price, stock, shipping time, return policy, reviews) and accept a revenue split.

If the agent's ranking logic is primarily governed by semantic fit and expected satisfaction, such a seller can outrank a global brand without any dedicated ad budget → simply by being the best answer to the user's request.

In economic terms, revshare removes the capital requirement for entry into the recommendation layer.

6.3. Incentives for reasoning quality

Revshare also changes the platform’s objective function:

Under PPC, a platform is incentivised to maximise clicks and bids, subject to not alienating users too much.

Under revshare, a platform is incentivised to maximise successful outcomes per unit of user time.

For an LLM-based agent, this means:

Asking clarifying questions rather than prematurely proposing options.

Avoiding deceptive or low-quality offers which lead to returns or cancellations.

Learning, via reinforcement, what patterns of reasoning lead to higher long-term satisfaction and revenue.

Stochastic reasoning is no longer a cost centre; it becomes the core asset being monetised.

7. OpenAI’s burn as economic investment

7.1. Training as capital expenditure

If we treat large-scale training runs as a form of capital expenditure (CapEx) on “intelligence infrastructure”, then:

The up-front billions spent on training are analogous to building a global logistics network or a hyperscale data centre fleet.

Once built, these models can be amortised over many years of inference and fine-tuning.

Epoch’s analysis suggests around US$3 billion of training compute in 2024, with additional research compute amortised over multiple years (Epoch, 2025b).

In this framing, OpenAI’s apparently extreme burn is not mere consumption; it is acquiring a durable asset → a global reasoning engine that, once sufficiently efficient, can serve trillions of tokens at low marginal cost.

7.2. Inference as variable cost with declining unit price

Inference is closer to a variable cost: each token generated consumes compute and thus incurs cost. However, several trends push unit cost down over time:

Hardware improvements (GPUs, TPUs, specialised accelerators).

Software optimisation (better kernels, quantisation, sparsity, batching).

Model architecture improvements (more efficient transformers, Mixture-of-Experts, distillation).

The combination means that even as total inference spend rises with usage, cost per token falls, improving the economics of revshare.

7.3. Revshare as the natural monetisation of this asset

Given this cost structure, revshare is appealing because:

It scales with value delivered, not tokens consumed.

It provides a natural hedge against compute costs: if compute is expensive, prices to merchants and users can be adjusted; if compute becomes cheap, margins expand.

It aligns the long-term interests of OpenAI (or any LLM platform), merchants, and users.

In other words, revshare turns “burn” into a leveraged call option on the future of AI-mediated commerce.

8. Forecast: from ads to agentic marketplaces

8.1. Macro context

Global advertising spend is expected to exceed US$1.0–1.1 trillion in 2025, with digital formats taking an ever-larger share (WARC, 2025; eMarketer, 2025). Digital advertising and marketing could reach US$1.5 trillion by 2030 according to some pre-AI forecasts (GlobeNewswire, 2023).

At the same time, entertainment and media are projected to grow to roughly US$3.5 trillion by 2029, as new modes of value creation emerge (PwC, 2025).

In this environment, we can imagine three layers:

Legacy ad-driven discovery (search, social).

AI-augmented ad tools (better targeting, creative optimisation).

Agentic marketplaces where autonomous or semi-autonomous agents mediate a growing share of transactions directly.

Revshare-driven LLM agents sit in layer 3.

8.2. Substitution and expansion

If agentic marketplaces capture even 5–10% of global e-commerce and services by 2030, the implied GMV is enormous. With revshare rates around 10–15%, the revenue pool available to AI platforms could rival or exceed current PPC markets.

Crucially, this is not purely substitution:

Agents can create new demand by making complex purchases easier (e.g. multi-city travel, tradesperson coordination, multi-step restorations).

They can unlock latent long-tail supply by giving visibility to small, high-fit providers previously priced out of PPC.

The result is a combined substitution + expansion story.

9. Conclusion

Frontier LLMs are expensive to build and run. OpenAI's current economics reflect this clearly: billions of dollars of annual compute spend, partially offset by rapidly growing → -but still relatively narrow → -revenue streams (Epoch, 2025b; Financial Times, 2025).

At the same time, the search → ad model that funded the last era of the internet is showing its limits: rising CPCs, entrenched incumbency, and mis-aligned incentives that reward attention rather than outcomes (WordStream, 2025b; StoreGrowers, 2025).

This paper has argued that:

Stochastic reasoning makes LLMs economically incompatible with auction-driven ads, because such auctions distort the chain of thought that users rely on.

Revshare → revenue paid on completed outcomes → fits the architecture of LLMs and aligns incentives for users, merchants, and platforms.

A single high-intent conversation can plausibly generate several dollars of gross margin; at scale, this can fund frontier compute while preserving reasoning quality.

Revshare removes the capital requirement for visibility, giving small merchants a realistic chance to compete based on merit.

Over the medium term, agentic marketplaces are likely to capture a meaningful share of global digital commerce and services, positioning outcome-based revenue models as central to the next phase of internet economics.

In short:

Search monetises fragments of intent.

LLMs monetise the entire reasoning chain.

The only business model that respects this difference → and turns it into a sustainable economic engine → is revshare.

References

Cottier, M. et al. (2024) The rising costs of training frontier AI models, arXiv preprint.

Epoch (2025a) ‘OpenAI’s revenue has been growing 3x a year since 2024', Epoch AI, 14 October. Available at: https://epoch.ai/data-insights/openai-revenue

Epoch (2025b) ‘Most of OpenAI’s 2024 compute went to experiments’, Epoch AI, 10 October. Available at: https://epoch.ai/data-insights/openai-compute-spend

Entrepreneur (2025) ‘OpenAI saw more revenue in six months than all of last year’, Entrepreneur, 30 September. Available at: https://www.entrepreneur.com/business-news/openai-saw-more-revenue-in-six-months-than-all-of-last-year/497774

Financial Times (2025) ‘How high are OpenAI’s compute costs? Possibly a lot higher than previously thought’, Financial Times, 12 November. Available at: https://www.ft.com/content/fce77ba4-6231-4920-9e99-693a6c38e7d5

Globe Newswire (2023) ‘Digital Advertising and Marketing Global Market to Reach $1.5 Trillion by 2030’. Available at: https://www.globenewswire.com/news-release/2023/03/23/2633485/28124/en/Digital-Advertising-and-Marketing-Global-Market-to-Reach-1-5-Trillion-by-2030-Influencer-Marketing-is-Powerful-Weapon-for-Digital-Marketing-Teams.html

LocaliQ (2025) ‘Search Advertising Benchmarks 2025’, LocaliQ, 2 June. Available at: https://localiq.com/blog/search-advertising-benchmarks/

Marketing-Interactive (2025) ‘CMOs must adapt, as global ad spend becomes overwhelmingly digital’, 20 October. Available at: https://www.marketing-interactive.com/cmos-must-adapt-as-global-ad-spend-becomes-overwhelmingly-digital

PwC (2025) ‘Global Entertainment & Media Outlook 2025–2029’, PwC. Available at: https://www.pwc.com/gx/en/issues/business-model-reinvention/outlook/insights-and-perspectives.html

Reuters (2025) ‘OpenAI’s annualized revenue hits $10 billion, up from $5.5 billion in December 2024', 10 June. Available at: https://www.reuters.com/business/media-telecom/openais-annualized-revenue-hits-10-billion-up-55-billion-december-2024-2025-06-09/

StoreGrowers (2025) ’27 Google Ads Benchmarks (2025)’, StoreGrowers, 13 June. Available at: https://www.storegrowers.com/google-ads-benchmarks/

The Register (2025) ‘OpenAI has spent $12B on inference with Microsoft: Report’, The Register, 12 November. Available at: https://www.theregister.com/2025/11/12/openai_spending_report/

WARC (2025) ‘Global advertising spend to pass $1 trillion for the first time this year’, WARC. Available at: https://www.warc.com/content/feed/global-advertising-spend-to-pass-1-trillion-for-the-first-time-this-year/10119

WordStream (2025a) ‘Google Ads Benchmarks 2025: Competitive Data & Insights’, WordStream, 29 September. Available at: https://www.wordstream.com/blog/2025-google-ads-benchmarks

WordStream (2025b) ‘How much does Google Ads cost in 2025?’, WordStream, 24 October. Available at: https://www.wordstream.com/blog/google-ads-cost

eMarketer (2025) ‘Worldwide Ad Spending Forecast 2025’, eMarketer, 31 January. Available at: https://www.emarketer.com/content/worldwide-total-media-ad-spending-cross-1-trillion-this-year