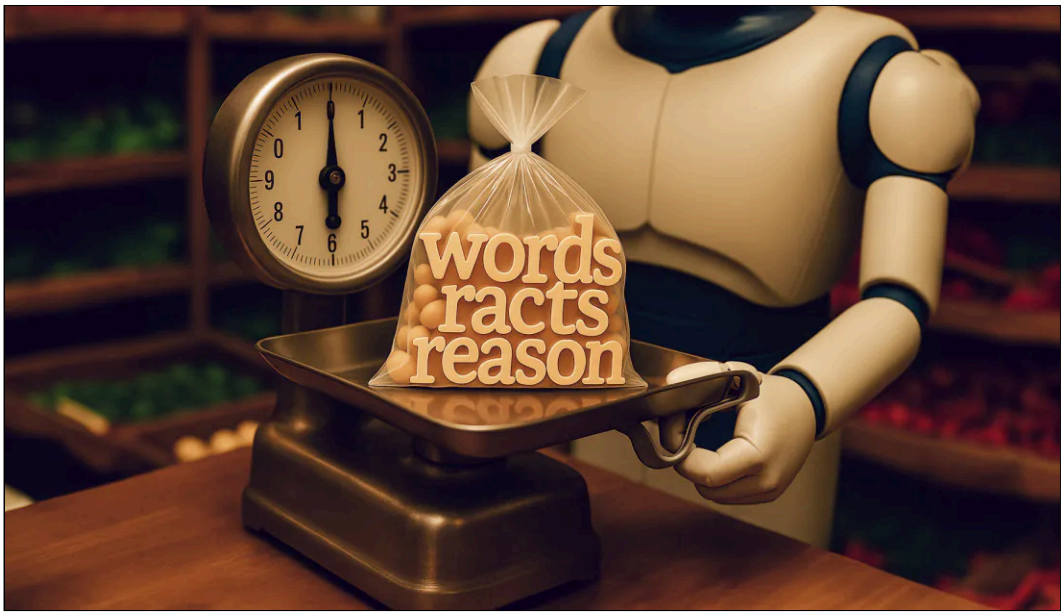
New Research Papers Question ‘Token’ Pricing for AI Chats

Originally published at https://www.unite.ai/new-research-papers-question-token-pricing-for-ai-chats/



New research shows that the way AI services bill by tokens hides the real cost from users. Providers can quietly inflate charges by fudging token counts or slipping in hidden steps. Some systems run extra processes that don’t affect the output but still show up on the bill. Auditing tools have been proposed, but without real oversight, users are left paying for more than they realize.

In nearly all cases, what we as consumers pay for AI-powered chat interfaces, such as ChatGPT-4o, is currently measured in tokens: invisible units of text that go unnoticed during use, yet are counted with exact precision for billing purposes; and though each exchange is priced by the number of tokens processed, the user has no direct way to confirm the count.

Despite our (at best) imperfect understanding of what we get for our purchased ‘token’ unit, token-based billing has become the standard approach across providers, resting on what may prove to be a precarious assumption of trust.

Token Words

A token is not quite the same as a word, though it often plays a similar role, and most providers use the term ‘token’ to describe small units of text such as words, punctuation marks, or word-fragments. The word *‘unbelievable’*, for example, might be counted as a single token by one system, while another might split it into *un*, *believ* and *able*, with each piece increasing the cost.

This system applies to both the text a user inputs and the model’s reply, with the price based on the total number of these units.

The difficulty lies in the fact that users *do not get to see this process*. Most interfaces do not show token counts while a conversation is happening, and the way tokens are calculated is hard to reproduce. Even if a count is shown *after* a reply, it is too late to tell whether it was fair, creating a mismatch between what the user sees and what they are paying for.

Recent research points to deeper problems: one study shows how providers can overcharge without ever breaking the rules, simply by inflating token counts in ways that the user cannot see; another reveals the mismatch between what interfaces display and what is actually billed, leaving users with the illusion of efficiency where there may be none; and a third exposes how models routinely generate internal reasoning steps that are never shown to the user, yet still appear on the invoice.

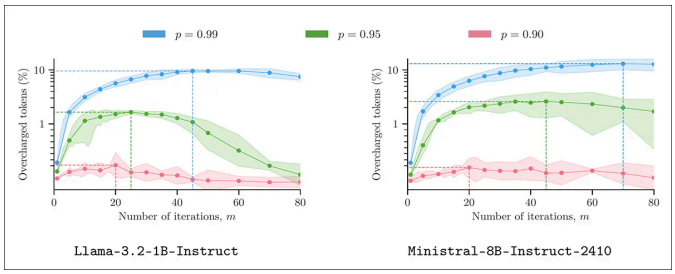
The findings depict a system that *seems* precise, with exact numbers implying clarity, yet whose underlying logic remains hidden. Whether this is by design, or a structural flaw, the result is the same: users pay for more than they can see, and often more than they expect.

Cheaper by the Dozen?

In the first of these papers – titled Is *Your LLM Overcharging You? Tokenization, Transparency, and Incentives*, from four researchers at the Max Planck Institute for Software Systems – the authors argue that the risks of token-based billing extend beyond opacity, pointing to a built-in incentive for providers to inflate token counts:

> *'The core of the problem lies in the fact that the tokenization of a string is not unique. For example, consider that the user submits the prompt "Where does the next NeurIPS take place?" to the provider, the provider feeds it into an LLM, and the model generates the output "|San| Diego|" consisting of two tokens.*
>
> *'Since the user is oblivious to the generative process, a self-serving provider has the capacity to misreport the tokenization of the output to the user without even changing the underlying string. For instance, the provider could simply share the tokenization "|S|a|n| |D|i|e|g|o|" and overcharge the user for nine tokens instead of two!'*

The paper presents a heuristic capable of performing this kind of disingenuous calculation without altering visible output, and without violating plausibility under typical decoding settings. Tested on models from the LLaMA, Mistral and Gemma series, using real prompts, the method achieves measurable overcharges without appearing anomalous:



Token inflation using 'plausible misreporting'. Each panel shows the percentage of overcharged tokens resulting from a provider applying Algorithm 1 to outputs from 400 LMSYS prompts, under varying sampling parameters (m and p). All outputs were generated at temperature 1.3, with five repetitions per setting to calculate 90% confidence intervals. Source: https://arxiv.org/pdf/2505.21627

To address the problem, the researchers call for billing based on *character count* rather than tokens, arguing that this is the only approach that gives providers a reason to report usage honestly, and contending that if the goal is fair pricing, then tying cost to visible characters, not hidden processes, is the only option that stands up to scrutiny. Character-based pricing, they argue, would remove the motive to misreport while also rewarding shorter, more efficient outputs.

Here there are a number of extra considerations, however (in most cases conceded by the authors). Firstly, the character-based scheme proposed introduces additional business logic that may favor the vendor over the consumer:

> *'[A] provider that never misreports has a clear incentive to generate the shortest possible output token sequence, and improve current tokenization algorithms such as BPE, so that they compress the output token sequence as much as possible'*

The optimistic motif here is that the vendor is thus encouraged to produce concise and more meaningful and valuable output. In practice, there are obviously less virtuous ways for a provider to reduce text-count.

Secondly, it is reasonable to assume, the authors state, that companies would likely require legislation in order to transit from the arcane token system to a clearer, text-based billing method. Down the line, an insurgent startup may decide to differentiate their product by launching it with this kind of pricing model; but anyone with a truly competitive product (and operating at a lower scale than EEE category) is disincentivized to do this.

Finally, larcenous algorithms such as the authors propose would come with their own computational cost; if the expense of calculating an 'upcharge' exceeded the potential profit benefit, the scheme would clearly have no merit. However the researchers emphasize that their proposed algorithm is effective and economical.

The authors provide the code for their theories at GitHub.

The Switch

The second paper – titled *Invisible Tokens, Visible Bills: The Urgent Need to Audit Hidden Operations in Opaque LLM Services*, from researchers at the University of Maryland and Berkeley – argues that misaligned incentives in commercial language model APIs are not limited to token splitting, but extend to *entire classes* of hidden operations.

These include internal model calls, speculative reasoning, tool usage, and multi-agent interactions – all of which may be billed to the user without visibility or recourse.

Provider	Visible?	Pricing
OpenAI o1 Jaech et al. [2024]	✗	$60 / MTok
OpenAI o3 Jaech et al. [2024]	✗	$40 / MTok
OpenAI o1-pro Jaech et al. [2024]	✗	$600 / MTok
Gemini 2.5 Pro Anil et al. [2023]	✗	$15 / MTok
Claude Opus 4 Anthropic [2025]	✗	$75 / MTok

Pricing and transparency of reasoning LLM APIs across major providers. All listed services charge users for hidden internal reasoning tokens, and none make these tokens visible at runtime. Costs vary significantly, with OpenAI's o1-pro model charging ten times more per million tokens than Claude Opus 4 or Gemini 2.5 Pro, despite equal opacity. Source: https://www.arxiv.org/pdf/2505.18471

Unlike conventional billing, where the quantity and quality of services are verifiable, the authors contend that today's LLM platforms operate under *structural opacity*: users are charged based on reported token and API usage, but have no means to confirm that these metrics reflect real or necessary work.

The paper identifies two key forms of manipulation: *quantity inflation*, where the number of tokens or calls is increased without user benefit; and *quality downgrade*, where lower-performing models or tools are silently used in place of premium components:

> *'In reasoning LLM APIs, providers often maintain multiple variants of the same model family, differing in capacity, training data, or optimization strategy (e.g., ChatGPT o1, o3). Model downgrade refers to the silent substitution of lower-cost models, which may introduce misalignment between expected and actual service quality.*
>
> *'For example, a prompt may be processed by a smaller-sized model, while billing remains unchanged. This practice is difficult for users to detect, as the final answer may still appear plausible for many tasks.'*

The paper documents instances where more than ninety percent of billed tokens were never shown to users, with internal reasoning inflating token usage by a factor greater than twenty. Justified or not, the opacity of these steps denies users any basis for evaluating their relevance or legitimacy.

In agentic systems, the opacity increases, as internal exchanges between AI agents can each incur charges without meaningfully affecting the final output:

> *'Beyond internal reasoning, agents communicate by exchanging prompts, summaries, and planning instructions. Each agent both interprets inputs from others and generates outputs to guide the workflow. These inter-agent messages may consume substantial tokens, which are often not directly visible to end users.*
>
> *'All tokens consumed during agent coordination, including generated prompts, responses, and tool-related instructions, are typically not surfaced to the user. When the agents themselves use reasoning models, billing becomes even more opaque'*

To confront these issues, the authors propose a layered auditing framework involving cryptographic proofs of internal activity, verifiable markers of model or tool identity, and independent oversight. The underlying concern, however, is structural: current LLM billing schemes depend on a persistent *asymmetry of information*, leaving users exposed to costs that they cannot verify or break down.

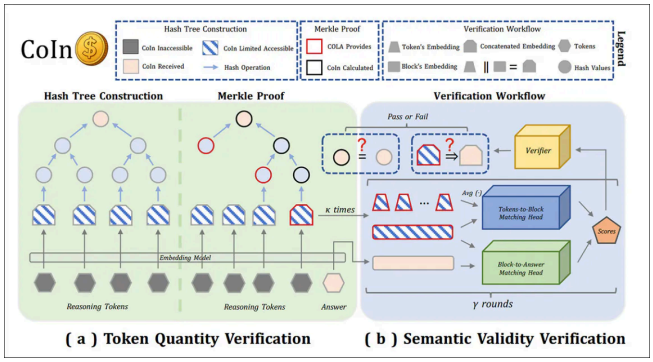
Counting the Invisible

The final paper, from researchers at the University of Maryland, re-frames the billing problem not as a question of misuse or misreporting, but of structure. The paper – titled *CoIn: Counting the Invisible Reasoning Tokens in Commercial Opaque LLM APIs*, and from ten researchers at the University of Maryland – observes that most commercial LLM services now hide the *intermediate reasoning* that contributes to a model's final answer, yet *still charge for those tokens*.

The paper asserts that this creates an unobservable billing surface where entire sequences can be fabricated, injected, or inflated without detection*:

> *'[This] invisibility allows providers to **misreport token counts** or **inject low-cost, fabricated reasoning tokens to artificially inflate token counts**. We refer to this practice as **token count inflation**.*
>
> *'For instance, a single high-efficiency ARC-AGI run by OpenAI's o3 model consumed 111 million tokens, costing $66,772.3 Given this scale, even small manipulations can lead to substantial financial impact.*
>
> *'Such information asymmetry allows AI companies to significantly overcharge users, thereby undermining their interests.'*

To counter this asymmetry, the authors propose *CoIn*, a third-party auditing system designed to verify hidden tokens without revealing their contents, and which uses hashed fingerprints and semantic checks to spot signs of inflation.



Overview of the CoIn auditing system for opaque commercial LLMs. Panel A shows how reasoning token embeddings are hashed into a Merkle tree for token count verification without revealing token contents. Panel B illustrates semantic validity checks, where lightweight neural networks compare reasoning blocks to the final answer. Together, these components allow third-party auditors to detect hidden token inflation while preserving the confidentiality of proprietary model behavior. Source: https://arxiv.org/pdf/2505.13778

One component verifies token counts cryptographically using a Merkle tree; the other assesses the relevance of the hidden content by comparing it to the answer embedding. This allows auditors to detect padding or irrelevance – signs that tokens are being inserted simply to hike up the bill.

When deployed in tests, CoIn achieved a detection success rate of nearly 95% for some forms of inflation, with minimal exposure of the underlying data. Though the system still depends on voluntary cooperation from providers, and has limited resolution in edge cases, its broader point is unmistakable: the very architecture of current LLM billing assumes an honesty that cannot be verified.

Conclusion

Besides the advantage of gaining pre-payment from users, a scrip-based currency (such as the 'buzz' system at CivitAI) helps to abstract users away from the true value of the currency they are spending, or the commodity they are buying. Likewise, giving a vendor leeway to define their own units of measurement further leaves the consumer in the dark about what they are actually spending, in terms of real money.

Like the lack of clocks in Las Vegas, measures of this kind are often aimed at making the consumer reckless or indifferent to cost.

The scarcely-understood *token*, which can be consumed and defined in so many ways, is perhaps not a suitable unit of measurement for LLM consumption – not least because it can cost many times more tokens to calculate a poorer LLM result in a non-English language, compared to an English-based session.

However, character-based output, as suggested by the Max Planck researchers, would likely favor more concise languages and penalize naturally verbose languages. Since visual indications such as a depreciating token counter would probably make us a little more spendthrift in our LLM sessions, it seems unlikely that such useful GUI additions are coming anytime soon – at least without legislative action.

** Authors' emphases. My conversion of the authors' inline citations to hyperlinks.*

First published Thursday, May 29, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More