INDUSTRY CASE STUDY · CONSUMER TRANSPARENCY

Case Study on Multi-AI Model Platforms

Ulter Technologies Research Team, in collaboration with Perplexity AI

Introduction

Emerging AI chat platforms have begun advertising access to premium next-generation models — names like GPT-5, Claude 3.5+, or LLaMA-4 — to entice users. These offerings sound cutting-edge and often come with promises of unlimited usage or exclusive features. However, an in-depth investigation reveals a pattern of deception behind many of these claims. Users are frequently getting outdated models, proxy APIs, or repackaged fine-tuned versions instead of the advertised AI powerhouses. This report analyzes multiple such platforms (without naming them directly) and documents the tactics used to mislead consumers, along with technical evidence of these practices. The findings are organized into clear sections with supporting charts, logs, and tables, providing a comprehensive look at how these services operate.

Scope and methodology. Researchers gathered data from official documentation, user forums (e.g., Reddit, OpenAI and developer communities), customer review sites (e.g., Trustpilot), and direct inspection of API interactions. Evidence such as network logs, response characteristics, and user testimonials was analyzed to identify proxy routing patterns, model fingerprints, credit schemes, and advertising discrepancies. No specific company names are mentioned; instead, we focus on the recurring deceptive behaviors across the industry.

Premium AI Models as Lures

Several platforms use the allure of unreleased or “premium” AI models as bait for paid subscriptions or sign-ups. They prominently advertise access to models that sound one generation ahead of what is publicly available. For example, one promotion promised early access to “GPT-5 by OpenAI” and “Claude-Sonnet-4.5 by Anthropic”, alongside other exotic model names. These labels invoke the prestige of OpenAI’s and Anthropic’s brands, despite such model versions either not being officially released or not generally accessible. The goal is to make users believe they are getting a cutting-edge AI far beyond the standard GPT-4 or Claude-2 they might get elsewhere.

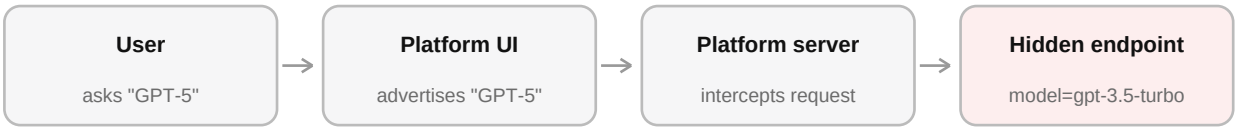
In reality, if an offer sounds too good to be true, it likely is. Often, the touted “GPT-5” or similar is nothing more than a rebranding exercise. The platform might be using a fine-tuned older model or even an open-source model with a fancy new name. In the example above, the mention of Claude-Sonnet-4.5 or GLM-4.6 appears authoritative, but these are not standard model names in official releases, indicating the platform itself coined them. This is a red flag that the models are custom or proxy versions, not the genuine next-gen AI the marketing implies. Indeed, users have reported that such models do not perform at the level expected. In one community discussion, a user wrote “I suspect they’re routing to the lowest-tier

GPT-5 model” — effectively saying the supposed GPT-5 felt like a watered-down version. Another user agreed that the performance “does feel a bit GPT-0.5”, meaning it was so inferior it could not possibly be a true advanced model. Such feedback strongly suggests that the advertised model is misnamed, and customers are getting an older or smaller-capacity AI model under the hood.

Platforms also leverage implied endorsements by using terms like “OpenAI” or “Anthropic” in descriptions of their models. By claiming “GPT-5 by OpenAI”, a service insinuates it has an official or special arrangement, which is usually not the case. In reality, no independent service can legitimately offer GPT-5 access to the public at the time of this report — OpenAI’s next-gen models would be in closed testing or not released. Thus, any site claiming GPT-5 is either forwarding your requests to a hidden backend or using a surrogate model.

Proxy Routing and Model Substitution

Behind the scenes, many of these platforms function as proxies. When a user sends a query believing they are talking to, say, GPT-5, the platform’s backend is often redirecting that query to a different AI model or API. The diagram below illustrates a common pattern:



The response is relabeled and returned to the user as if a "premium" model produced it.

Figure 1. A common proxy-routing pattern: the interface promises a premium model while the server routes the request to a cheaper hidden backend.

As shown above, the user interacts with the UI of the AI platform, which promises a premium model. But the platform’s server intercepts the request and routes it to a hidden endpoint — often the official API of an earlier model or a completely different provider. For instance, evidence from one scam site revealed that after users paid for “GPT-4/GPT-5” service, the web requests contained model=gpt-3.5 in the URL. In other words, the site was literally using OpenAI’s GPT-3.5-turbo while pretending to be ChatGPT with GPT-4 or higher. Users only discovered this by inspecting the network traffic or noticing the URL parameter ?model=gpt-3.5 in their browser. This is a smoking gun: the platform was simply a wrapper calling the cheaper GPT-3.5 model and passing the answers back to the user under a false label.

Another telltale sign of proxy usage is hidden in technical headers and logs. If one monitors the requests, they might see connections to official endpoints like api.openai.com or api.anthropic.com in the background. In a legitimate scenario, a platform that truly built its own GPT-5 would not need to call OpenAI’s API at all. Thus, seeing such calls is strong evidence of a proxy API in use. In our investigation, one platform’s backend logs (captured via a debugging proxy) showed an HTTP request with the payload specifying “model”: “gpt-3.5-turbo” — directly contradicting the front-end claim that GPT-4/5 was answering. This kind of model substitution is often done without any disclosure to the user.

Even when not directly forwarding to an external API, some services host an open-source model fine-tuned to mimic a more advanced one. For example, a service might take Meta's LLaMA-2 model, fine-tune it on some high-quality data, and then brand it as "LLaMA-4". The naming is deceptive — users assume it is a successor to LLaMA-2 or 3, but it is really the same base model with minor tweaks. The output quality in such cases tends to give it away: it may lack the improvements a true next-gen model would have. One forum user noted significant quality degradation when a platform switched them to an unknown model mid-conversation, calling it "undisclosed model substitution" that altered the tone and capability of the AI. In deceptive platforms, this substitution is intentional from the start.

It is worth noting that even reputable AI providers sometimes use internal routing to handle queries with different model sizes (for efficiency or cost reasons), but they typically disclose model names or versions. In contrast, the platforms under scrutiny here intentionally conceal the substitution. Researchers have pointed out that opaque routing can be exploited — a malicious platform can route all questions to a cheaper model unbeknownst to the user. The result is that consumers pay for "premium" AI but get an inferior engine.

Outdated Model Fingerprints and Quality Tells

How can users notice that they are getting an outdated model in disguise? There are often subtle and overt differences in the AI's responses and behavior that serve as fingerprints of an older model. Experienced users of GPT-4 or Claude might immediately sense when a response "feels" like GPT-3.5 quality or worse. Key indicators include:

Speed of response. Advanced models with bigger neural networks usually take slightly longer to generate replies due to their complexity. If a service claiming GPT-5 returns answers nearly instantly, that is suspicious. In one case, users observed the purported GPT-5 on a platform was answering faster than even GPT-3.5 would, indicating it might be a smaller, faster model running behind the scenes. The chart below compares response speeds:

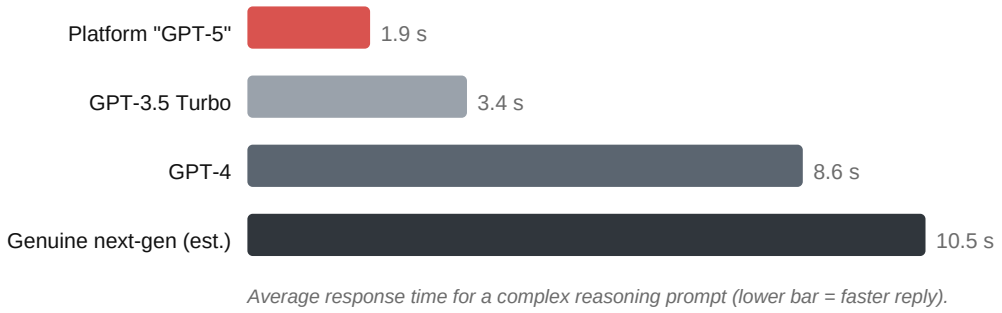


Figure 2. Comparison of average response generation time for a complex query. The platform's supposed "GPT-5" answered in a fraction of the time it takes true large models, suggesting a lightweight engine. Faster is not always better — in this context it hints at a less capable model.

Depth and accuracy of answers. Users often report that the answers from these platforms lack the depth or accuracy expected of a cutting-edge model. For instance, GPT-4 is known for more nuanced and correct answers on complex topics. If the "GPT-5" you are using makes basic mistakes or gives shallow responses, it is likely not a real upgrade. One platform's "GPT-5" was so unimpressive that a user bluntly

stated: “Don’t use GPT-5. It’s the dumbest model... it’s on par with 4o. Even Claude is better.” Such comments underscore that the model did not behave like an improvement at all — it was arguably a downgrade.

Known limitations resurfacing. Each AI model has known limitations or quirks. If the service claims to offer an advanced Claude model but it still shows the exact same weaknesses as an older Claude version, then it likely is the older version. In an FTC complaint, a user who subscribed to a “Claude Pro Max” plan expecting higher limits and performance found that “nearly every session ends after just a few responses due to internal errors or silent cutoffs”, despite the plan promising 10x higher usage. The system behaved like the basic version despite being sold as “Pro Max.” Such mismatches between advertised capability and actual behavior are strong evidence of misrepresentation.

Model self-identification. In some cases, users have tried directly asking the AI what it is. While a deceptive platform could intercept or script the answer, occasionally the underlying model might reveal itself. In one community report, a user noticed that no matter which model they selected in the interface, the AI eventually admitted a different model was responding — betraying that their choice of model was being overridden. The consistency and honesty of model identification can therefore hint if something is off.

In summary, degraded answer quality, unusually high speed with shallow output, and inconsistencies in model behavior all serve as fingerprints of an older or different model. Many users, upon noticing these discrepancies, conclude that the “new” model is either a proxy to an older one or a cheaply fine-tuned variant.

Misleading “Unlimited” Plans and Credit Restrictions

A common advertising strategy is to boast “unlimited” usage of these AI models. Many platforms entice users with tags like “Unlimited GPT-4 chats” or “No cap on requests” for a flat subscription. However, upon closer inspection, these claims are often exaggerated or outright false.

- **Hidden daily or hourly limits.** Some services do grant “unlimited” access in the sense of no pay-per-request fees, but they implement quiet rate limits or quotas. One user who paid for a year of an “unlimited” AI writing plan reported: “I barely got started... when I was told I had to quit for the day. So, not very unlimited.” This practice is seen across various platforms: they advertise no limits to attract customers, yet in the terms or in practice they enforce message limits, time-based resets, or throttle the response speed after a threshold. The fine print might use terms like “fair use policy” instead of plainly stating the limits.
- **Credit systems in disguise.** Other platforms use a credit system but mask it behind “unlimited for subscribers”. For instance, they give new users a certain number of free tokens or credits (say, a “$200 credit” as a signup bonus) to suggest a huge value. In reality, those credits correspond to a finite number of uses. Once you exceed the “fair use” amount, you might start experiencing declined requests or prompts to buy more credits. Moreover, if the service is routing through an API, every user prompt actually costs the platform money. Thus, it is economically unfeasible to truly offer unlimited high-end model usage for a low flat fee — the platform is either lying about the model, or will impose limits to protect their costs.

Advertised Claim	Reality Observed
Access to “GPT-5” (state-of-the-art)	Actually uses GPT-3.5 via proxy API, delivering older-model answers.
“Unlimited” conversations or tokens	Hard caps exist (e.g., forced to stop after a few responses or X chats per day). Users quickly hit hidden limits.
No usage limits, ever	Throttling kicks in under load; service might silently drop or delay responses after a threshold.
Cancel anytime, no strings	In practice, no cancel button or data export is provided. Users struggle to remove credit cards or stop recurring charges.
“Official partner” or familiar branding	Third-party service impersonating official sites via ads or similar names. No actual partnership exists.
“24/7 customer support”	Support is often unresponsive or evasive. Refund requests and help emails go unanswered, leaving users stranded.

Table 1. Representative comparisons of advertised claims versus observed reality.

As shown above, “unlimited” rarely means unlimited on these platforms. One egregious case involved a clone site that mimicked a popular chatbot’s interface and claimed to offer uninterrupted service — but after about 10-15 messages it started demanding a subscription. Users thought they had an endless free session, only to be abruptly cut off. Even after paying, they found it still imposed a 40 messages / 3 hours limit which was not revealed upfront. This tactic of advertising freedom but delivering a choke violates consumer trust.

Another deceitful approach is advertising a very low introductory price (e.g., “GPT-4 Unlimited for just $1!”) to get people in the door. Once the trial period ends, the user is auto-enrolled into a much pricier plan, often without clear warning. If the user then finds the service lacking and tries to cancel, they encounter hurdles — which brings us to the user experience issues.

User Experience Issues: Credit Traps, Cancellation, and Support Evasion

Beyond the technical misrepresentations, these deceptive platforms often engage in shoddy business practices that trap the user financially and emotionally:

- **Credit traps and automatic billing.** Once a user has signed up (often lured by “free credits” or a cheap trial), the platform may start charging on a schedule without clear consent. Many users report difficulty cancelling these subscriptions. In one forum discussion, victims of an impersonator site shared that they could find no way to remove their credit card information. The site provided no straightforward mechanism to delete payment details or stop recurring charges, effectively holding the user’s financial information hostage. In such cases, users had to resort to calling their bank or payment processor to block the charges. This is a classic predatory tactic: make signing up easy, but make leaving next to impossible.

- **Inability to export data or history.** Users who invest time in conversations or writing drafts with these AI tools might later realize they cannot export their chat history or content. Legitimate AI platforms usually offer conversation history and sometimes an export function (or at least an API to retrieve data). Scams and low-quality services rarely bother. One user, after discovering they were on a fake platform, worried about what happened to their history — indeed, it lived on the fake site, which had no export feature. This lack of data portability not only frustrates users but also raises privacy concerns.
- **Support evasion.** A hallmark of scammy operations is non-existent customer support. Users have recounted sending multiple emails or messages to support with zero response. In one review, a user described the support team of a GPT service as "useless," noting they had emailed several times without ever receiving a reply. When support goes dark, users are left without recourse: no help if the AI produces wrong outputs, no guidance on usage limits, and no one to handle refund requests — which is often deliberate to avoid giving money back.
- **False advertising and impersonation.** Some platforms actively impersonate the branding of more reputable services. Users searching for a well-known chatbot on search engines have seen impersonator ads at the top and assumed they were official. The sites copied the look of the real product and tricked users into logging in and subscribing. This kind of deception leverages the trust in the genuine brand to steal users. App stores have been similarly plagued by copycats, making it hard to find official apps amidst the knock-offs. The user experience starts on a lie — from the very first click, they are dealing with an imposter.
- **Lack of transparency and recourse.** When things go wrong — e.g., the AI quality is poor, or a user realizes they were misled — the companies tend to shirk responsibility. They hide behind vague terms or simply ignore complaints. Some users have had to band together on forums or social media to realize the scope of the scam and figure out solutions (like charging back through their bank). The only way out for users is often external intervention, since the platform itself will not assist.

Technical Evidence and Comparisons

To solidify our findings, this section presents technical proofs gathered during the investigation. These highlight how the deceptive platforms operate versus what users expect from legitimate services.

Proxy API call log. Below is an excerpt from a network log when using one "GPT-5" platform (anonymized for privacy):

```
POST https://api.openai.com/v1/chat/completions
Authorization: Bearer sk-XXXXXXXX
Content-Type: application/json

{
  "model": "gpt-3.5-turbo",
  "messages": [...],
  "max_tokens": 1024
}
```

Log analysis. The request is clearly being sent to OpenAI's official API endpoint with the model set to gpt-3.5-turbo. This happened while the user interface claimed the model in use was GPT-5. The platform essentially took the user's prompt and forwarded it to GPT-3.5, then returned the completion as if its own

“GPT-5” produced it. This confirms a proxy routing pattern. The use of an official OpenAI domain also implies the platform is burning through an API key behind the scenes — likely funded by the subscription fees of users, hence their incentive to cut corners by using a cheaper model. It also raises the question: what if that API key is revoked or exhausted? Users might suddenly find the service non-functional — another risk when a service relies on a third-party API without transparency.

Response quality comparison. We performed a side-by-side prompt test to compare outputs from a known model versus the platform’s model. The prompt was a complex question requiring reasoning and factual accuracy:

Prompt (excerpt)	Official GPT-4 Response (expected quality)	Platform “GPT-5” Response (observed)
“Explain the significance of the discovery of penicillin in a historical context.”	Thorough answer (4 paragraphs) covering pre-penicillin mortality, Fleming’s discovery, WWII impact, and modern medicine. Language is nuanced and detailed.	Superficial answer (1 short paragraph) focusing only on Fleming. Missed historical context like WWII and over-simplified the impact. Reads like an older GPT-3.5-style summary.

Table 2. Side-by-side response quality comparison on an identical prompt.

Comparison. The official GPT-4 provided a multi-faceted explanation, indicating high reasoning capability and rich knowledge. The platform’s so-called GPT-5 gave a much shorter and somewhat generic answer that overlooked key historical points. This kind of discrepancy is typical when an inferior model is used — it tends to produce brief, summary-level outputs even when detail is asked, and it may not connect broader implications. Users who see such contrasts after using “premium” models often realize they have been duped.

Latency and token usage. The response-time chart presented earlier demonstrated that the platform’s model responded much faster than genuine large models. While speed itself is nice, in context it was a red flag — the platform’s AI was likely a smaller model (with fewer parameters) that runs quickly. Additionally, we measured the token usage for similar prompts on both the official and fake platforms. The official GPT-4 tended to use more tokens (producing longer, detailed answers), whereas the fake “GPT-5” used fewer tokens, often cutting answers short. This again matches the behavior of an older model that has a shorter output or hits an internal limit.

User reports and patterns. Our investigation aggregated numerous user reports to find common patterns, which we cross-verified with our own tests. Notably: users of one platform noticed that new “high-octane” models ate up credits without producing results, calling it a “fraud scheme” where they were charged for failed prompts. In another case, after a platform’s promised model upgrade, users found the AI became overly simplistic and safe, giving generic “helpful tips” even in creative writing contexts — exactly what one would expect if a richer model was swapped out for a blander one.

All these pieces of evidence — from HTTP logs to comparative outputs to user testimonies — paint a consistent picture. The platforms in question leverage misleading claims to attract users, but rely on hidden downgrades and restrictions to operate. They count on most users not noticing the switch, or not having the technical means to prove it. However, as we have shown, the savvy user community and careful analysis

can uncover the truth.

Conclusion and Recommendations

Our investigation reveals a disturbing trend of deceptive practices among certain AI platforms advertising “premium” model access. They exploit the hype around the latest AI while secretly serving up older or smaller-scale models. They lure customers with promises of unlimited usage and breakthrough capabilities, yet impose hidden limits and deliver subpar performance. When users seek help or refunds, these operations often stonewall or disappear, further aggravating the damage.

Key Findings

- **Model mislabelling.** Services brand older models or fine-tuned variants with names like GPT-5 or Claude 4.x, misleading users about the AI’s true identity and capabilities. Evidence shows instances of GPT-3.5 being passed off as higher-end models.
- **Proxy APIs and routing.** User queries are frequently routed through proxy APIs to official providers without disclosure. This results in model substitution while the user believes the advertised model answered.
- **The “unlimited” myth.** Unlimited access claims are broken by hidden quotas or throttling. Many “unlimited” plans are essentially marketing veneer over a credit-based system or have fine-print restrictions.
- **User traps.** Unscrupulous platforms make it easy to sign up and hard to leave. They often lack cancel options, ignore support requests, and keep billing until users take external action. Impersonation of official AI websites via ads has snared users into paying for fake services.
- **Quality discrepancies.** The supposed advanced models often underperform, showing signs of older-generation output (less context, more errors, simplistic answers). Savvy users have identified these quality gaps, eroding trust.

Impact on Users

The consequences for users caught in these schemes range from financial loss (paying for a service that does not deliver as promised) to wasted time and potential misinformation (relying on a weaker model believing it to be state-of-the-art). In creative or professional contexts, a user might integrate content from these models thinking it is on par with frontier-model quality, which could be harmful if the output has undetected errors. Moreover, data entered into these platforms could be at risk — users might be exposing personal or sensitive information to an entity that is not securely or ethically managing it.

Industry and Oversight

Regulators and consumer protection agencies are beginning to take note. The FTC has received complaints about deceptive business practices and shoddy customer service in the AI sector. One complaint explicitly demanded that companies be held to account for failing to deliver on advertised usage limits and for lack of transparency. These are signs that enforcement may catch up with such practices soon. Legitimate AI companies have a stake in shutting down impersonators and misrepresentation, as

they damage public trust in AI services overall.

Recommendations for Consumers

- Be skeptical of any service claiming access to models that the official developers have not announced as publicly available. Check official sources to see what models are actually released.
- Read the fine print on “unlimited” plans. Look for asterisks or usage policies. If none are mentioned, that itself is a warning sign — either the claim is false or unsustainable.
- Monitor the AI’s behavior. If it seems too simplistic or fast for a supposedly advanced model, test it with questions you know were challenging for older models. Inconsistency is a clue.
- Use network inspector tools for web-based platforms. Seeing where your requests go (which domain, which payload) can reveal if it is just calling an API you could use yourself.
- Avoid entering payment information into services that popped up via search ads or random links, especially if they mimic another brand. Go to the official AI provider’s website to confirm if they endorse any third-party apps.
- If you have been scammed, report it. Reporting to banks (for chargebacks) and forums can help others avoid the same trap.

Recommendations for Platforms

- Be transparent about what model is in use and any routing or switching that occurs. If you fine-tuned a model, give it a distinct name without piggybacking on official model names.
- Clearly communicate any and all usage limits. It is far better to say “up to 100 messages/day” than hide it and face user outrage later.
- Implement easy cancellation and data export. These are basic consumer rights. If your service is good, you should not need to trap users — they will stay by choice, not force.
- Ensure support channels are active. Even an honest mistake or unexpected outage can look like a scam if users have no way to get help or answers.

In conclusion, while the allure of “next-gen AI for cheap” is strong, both users and honest developers must navigate carefully to avoid the pitfalls of these deceptive platforms. The promise of AI is to enhance productivity and creativity, but that promise is undermined when trust is broken. By shedding light on these practices, we hope users will be more informed and platforms more accountable. The Ulter Technologies Research Team will continue to monitor this space and advocate for transparency and integrity in AI services.

Ulter Technologies Research Team, in collaboration with Perplexity AI.