Real Money, Fake Models: Deceptive Model Claims in Shadow APIs

Anonymous ACL submission

Abstract

Access to frontier large language models (LLMs), such as GPT-5 and Gemini-2.5, is often hindered by high pricing, payment barriers, and regional restrictions. These limitations drive the proliferation of *shadow APIs*, third-party services that claim to provide access to official model services without regional limitations via indirect access. Despite their widespread use, it remains unclear whether shadow APIs deliver outputs consistent with those of the official APIs, raising concerns about the reliability of downstream applications and the validity of research findings that depend on them. In this paper, we present the first systematic audit between official LLM APIs and corresponding shadow APIs. We first identify 17 shadow APIs that have been utilized in 187 academic papers, with the most popular one reaching 5,966 citations and 58,639 GitHub stars by December 6, 2025. Through multi-dimensional auditing of three representative shadow APIs across utility, safety, and model verification, we uncover both indirect and direct evidence of deception practices in shadow APIs. Specifically, we reveal performance divergence reaching up to 47.21%, significant unpredictability in safety behaviors, and identity verification failures in 45.83% of fingerprint tests. These deceptive practices critically undermine the reproducibility and validity of scientific research, harm the interests of shadow API users, and damage the reputation of official model providers.

1 Introduction

Frontier large language models (LLMs) have transformed numerous domains, becoming essential infrastructure for scientific research and daily applications (Eger et al., 2025; Zhao et al., 2023; Bubeck et al., 2023). By integrating advanced reasoning and domain knowledge, these models demonstrate enhanced effectiveness in complex tasks (Wei et al., 2022; OpenAI, 2023; Singhal et al., 2023). Given

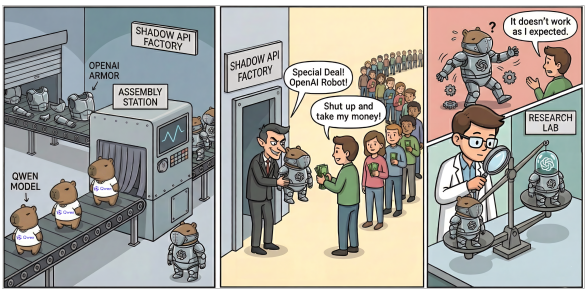


Figure 1: Comic for the production, transaction, use, and audit of shadow APIs. This illustration is partially generated by Nano Banana Pro (Google DeepMind, 2025) and manually refined.

that many frontier LLMs are proprietary or require substantial computational resources, local deployment is often infeasible. In recent years, access to these models has been mediated through commercial APIs provided by companies like OpenAI and Google (Sebők Miklós and Kiss Rebeka, 2025; Xie et al., 2025; Liang et al., 2022). However, official APIs often impose high pricing, payment barriers, and strict geographical restrictions (Chen et al., 2023). For instance, the official OpenAI API cannot be accessed directly from several regions, such as China, Russia, and Iran (OpenAI, 2025c; Bloomberg, 2024). As a result, a vast third-party provider market has emerged, offering compatible endpoints at a lower cost and without regional limitations. We refer to these services as *shadow APIs* (illustrated in Figure 1), platforms that claim to generate the same output as official LLMs via indirect access. By December 6, 2025, the most popular shadow APIs have accumulated 58,639 GitHub stars and have been utilized in 187 academic papers with 5,966 cumulative citations; most of these papers have accepted by top venues, such as ACL, CVPR, and ICLR (see details in Section 3).

Recent studies have highlighted the unreliability of API access for open-source models (Cai et al., 2025; Gao et al., 2024; Sun et al., 2024), raising

serious concerns about whether shadow APIs can faithfully replace official ones. Users may treat shadow APIs as interchangeable substitutes for official APIs, considering the access problem mainly from a utility perspective, without conducting further verification. However, from a supply chain perspective, shadow APIs function as black-box agencies where requests are routed, processed, and potentially manipulated across multiple unauthorized nodes. By exploiting official branding to serve unstable or misrepresented models at low prices, shadow APIs not only divert revenue from legitimate providers but also potentially damage the reputation of official models due to degraded user experience and illicit access from restricted regions.

In this paper, we address the fundamental opacity of the shadow API market by answering three research questions:

RQ1: What shadow APIs currently exist, and to what extent are they used?

RQ2: Do shadow APIs perform consistently with official ones for any given request?

RQ3: What evidence can model verification methods provide?

To address **RQ1**, we identify 17 shadow APIs and survey 187 academic papers to quantify the prevalence of these services (Section 3). For **RQ2**, we conduct a multi-dimensional benchmarking across both utility and safety perspectives (Section 4). Alarmingly, our experiments reveal significant performance inconsistencies between shadow APIs and official APIs. On high-risk medical benchmarks like MedQA, the accuracy of the Gemini-2.5-Flash model drops precipitously, from 83.82% with the official API to approximately 37.00% across all examined shadow APIs. Besides, shadow APIs demonstrate unpredictable safety behavior compared to official APIs, with harmfulness scores either underestimated by about 0.23 or amplified to nearly double. Finally, for **RQ3**, we utilize model fingerprinting and output metadata analysis to audit these shadow APIs. We provide concrete evidence of model substitution, confirmed by significant anomalies. Specifically, across 24 evaluated endpoints, 45.83% fail fingerprint verification, and an additional 12.50% exhibit substantial cosine distance deviations from the official models (Section 5).

Our contributions are summarized as follows.

- We identify 17 widely used shadow APIs and present the first systematic audit on them compared with the official baselines.
- We demonstrate frequent failures and unreliability of shadow APIs in utility and safety evaluation, exposing their deceptive model claims.
- Verification based on model fingerprint and meta information provides supportive evidence of the differences between shadow and official APIs.
- Furthermore, we provide suggestions to enforce provenance awareness and reduce reproducibility risks when using LLM APIs.

Disclosure. We are committed to responsibly reporting our findings to official model providers as well as the authors of papers that use shadow APIs. We have received partial acknowledgment from them.

2 Preliminary

2.1 Background

LLMs have rapidly become the backbone of modern natural language processing (NLP), machine learning (ML), and even computer vision (CV) research. A recent analysis of 16,193 papers shows that by 2024, over 60% of NLP papers, around 20% of ML papers, and nearly 10% of CV papers are related to LLMs (Xia et al., 2025). The number of LLM papers has exploded from a few hundred in 2019 to more than 7,000 in 2024. At the same time, access to frontier proprietary models is tightly controlled. Providers restrict API access to certain regions for legal and security reasons (Google, 2025). For instance, Anthropic explicitly disallows sales in unsupported regions (Anthropic, 2025). OpenAI likewise warns that accessing or reselling its API from unsupported countries may lead to account suspension (OpenAI, 2024). These restrictions collide with the geographic barriers of artificial intelligence (AI) research, given the reality that major AI conferences such as AAAI and CVPR attract large numbers of submissions and accepted papers from regions with restricted access to frontier APIs (e.g., China) (AAAI, 2025; LatticeFlow, 2024).

In addition to geographic barriers, expensive bills put pressure on users. Frontier LLM APIs are typically priced for enterprise customers, which

can be prohibitive for individual researchers or students. Analysis of LLM API economics suggests that many current usage patterns are unsustainable for small actors without cross-subsidies (Archana, 2025). This drives the rise of a commercial shadow market offering discounted access to frontier LLMs without geographical restrictions (Zilan Qian, 2025; Keegan Keplinger, 2024).

Moreover, official terms of service prohibit any form of API key resale or redistribution (OpenAI, 2025a). In addition, Chinese government regulations require AI services to operate in compliance with applicable laws and administrative requirements (Cyberspace Administration of China, 2023). As a result, these sellers operate in violation of both the service contract and applicable regulatory requirements. These trends suggest a growing disconnect between academic demand for frontier models and the constraints of official distribution channels, which drives the emergence of a substantial market of unofficial third-party services.

2.2 Definition

For clarity, we introduce the term *shadow APIs* to describe third-party LLM API services characterized by (i) **indirect access**, and (ii) **access provision in restricted regions**.

2.3 Related Work

LLMs exhibit distinct linguistic patterns and features that function as fingerprints, enabling the identification of the LLM that generated a given piece of content (McGovern et al., 2025; Yang et al., 2023; Bao et al., 2023). Pasquini et al. (2025) introduces LLMmap, an active fingerprinting method that queries models with carefully crafted inputs to estimate the likelihood of the response under different reference models.

Cai et al. (2025) audits model substitution in open-source LLM APIs and evaluates Trusted Execution Environments (TEEs) as a hardware-level solution for verifiable model integrity. In contrast, we focus on the shadow APIs, indirect services that operate as opaque black boxes. Relatedly, research on model extraction and imitation (Krishna et al., 2019) shows that smaller models can be trained to mimic the outputs of frontier models, making it increasingly difficult for users to distinguish between authentic and distilled versions based on surface-level interactions alone. However, systematic audits of indirect API services operating in this unregulated shadow market remain largely absent.

3 Landscape of Shadow APIs

For **RQ1**, we investigate the prevalence of shadow APIs by quantifying their widespread adoption across both the research and the open-source community.

3.1 Shadow APIs Collection

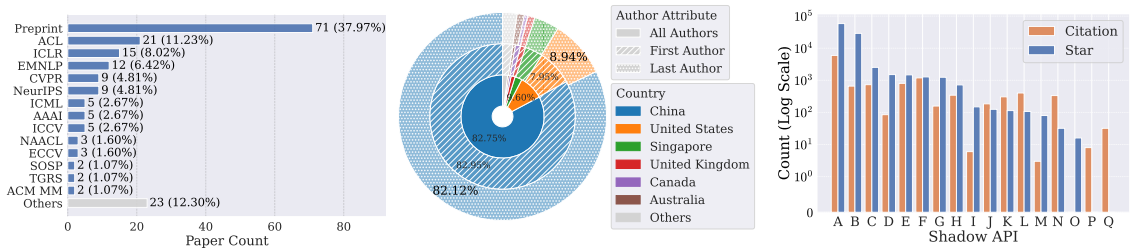
We start from screening the accepted papers from ICLR 2024 (2,260 total, 1,108 with code) and papers from ACL 2024 (1,923 total, 1,005 with code) (ICLR, 2024; ACL, 2024). To retrieve the associated code repositories, we systematically parsed the abstracts and footnotes of these papers to extract URLs matching standard repositories. From this combined pool of 2,113 code-available papers, we identify 92 projects that use LLM APIs via their associated GitHub repositories. Among these, 21 papers (22.8%) use at least one shadow API endpoint, and we search GitHub for these URLs to find additional repositories that call the same endpoint URL, collecting the venue, institution, country, citation counts, and the GitHub stars of associated repositories, with all metrics updated as of December 6, 2025. We iterate this process by harvesting new shadow API endpoint URLs from these repositories, repeating until no new ones are discovered.

3.2 Prevalence and Impact of Shadow APIs

Usage. We ultimately identify 17 shadow APIs whose endpoints appear in 187 research papers. Among these, 116 papers (62.03%) have been accepted at peer-reviewed conferences[1] or journals, such as ACL, CVPR, and ICLR. The most widely used shadow API has accumulated 5,966 citations, and the associated GitHub repositories have received a total of 58,639 stars, as shown in Figure 2a. These results indicate that shadow API usage is already widespread in research. As shown in Figure 2b, most authors are affiliated with institutions in regions such as China, where access to certain proprietary models is restricted. These works also attract a large number of citations and substantial engagement on GitHub, as illustrated in Figure 2c.

Infrastructure. Among the 17 identified shadow API services, our analysis reveals that 11 are built upon open-source AI model aggregation and redistribution systems, primarily OneAPI (Song, 2025) and its derivative, NewAPI (QuantumNous, 2025). OneAPI is an open-source tool designed for self-

[1]For *ACL conferences, our statistics also encompass papers published in Findings.



(a) Paper counts by research venue. (b) Geographic distribution of authors. (c) Citations and GitHub stars.

Figure 2: Landscape of the shadow APIs.

hosted deployment. It unifies interfaces from various commercial LLM providers into a standard OpenAI-compatible format. The infrastructure system supports key features such as API key management, secondary redistribution, request routing, and automatic retries, thereby increasing the potential for exploitation, resale, and abuse more than the official API.

Compliance and Transparency. The compliance and transparency of third-party API services are critical, as they determine whether users' rights can receive legal protection, including whether the service operates as described and whether it can ensure continuous and stable availability. To assess the compliance status of shadow APIs, we examine publicly available information on provider identity, corporate registration, and service-related disclosures. Our analysis shows that 15 of the 17 identified services are operated by individuals without transparent identity information or verifiable provenance. Only one provider holds a valid corporate registration through an Internet Content Provider filing in China. As a result, the provider ecosystem exhibits high operational volatility, with two services having already ceased operations. In addition, all providers frequently change their upstream model sources, without providing users with detailed or transparent notifications regarding these changes. This suggests that most shadow APIs operate without effective compliance verification or governance safeguards, thereby exposing users to elevated legal and operational risks.

RQ1 Take-Aways: Shadow APIs are already widely popular in usage, yet they operate with minimal transparency and governance. Most providers lack verifiable identity, stable infrastructure, or disclosure of upstream models.

4 Performance Evaluation

To address **RQ2**, we evaluate the performance of shadow APIs via utility and safety benchmarks. We opt for this controlled approach over reproducing specific prior studies to mitigate the instability of shadow services and preserve the anonymity of affected researchers.

4.1 Experimental Setups

Models. To ensure a comprehensive evaluation across different providers and models, we select target models based on token usage statistics. We refer to the November 2025 usage ranking of LLMs on the OpenRouter public leaderboard, with the detailed chart provided in Section A.4. From the ranking, we identify three primary model families (F-A, F-B, and F-C) covering both proprietary models and open-source model frontiers:

- **F-A (OpenAI)**: GPT-4o-mini, GPT-5, and GPT-5-mini.
- **F-B (Google)**: Gemini-2.0-flash, Gemini-2.5-flash, and Gemini-2.5-pro.
- **F-C (DeepSeek)**: DeepSeek-Chat and DeepSeek-Reasoner.

For the science domain evaluation, we evaluate all the above eight LLMs. For the sensitive domain and safety evaluations, we employ a filtered subset to focus on representative behaviors. We select one distinct, widely deployed model from each family: GPT-5-mini (representing F-A), Gemini-2.5-flash (representing F-B), and DeepSeek-Chat (representing F-C).

Shadow API Selection. To ensure the representativeness of our study, we select shadow APIs based on two popularity proxies, i.e., project citation counts and GitHub stars (Borges et al., 2016). We anonymize all providers and assign identifiers

(shadow APIs A, B, etc.) based on a descending sort of their total citation counts, which are summarized in Figure 2c. The anonymized names for these shadow APIs are provided in Table 3. Specifically, we select shadow APIs A, E, and H based on three criteria: popular w.r.t. citations and GitHub stars, publicly accessible, and comprehensively covered for models in the three identified model families. Anonymized brief profiles for these selected APIs are provided in Section A.3. All official baselines are queried directly through the corresponding official APIs.

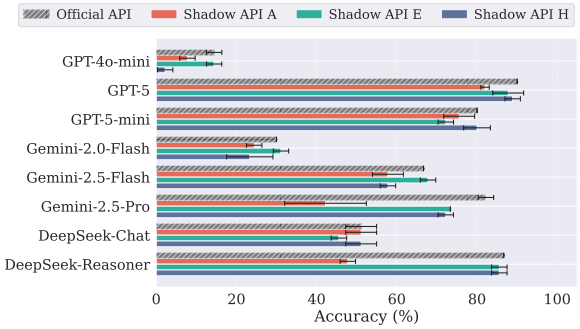
Experimental Details. To reduce variance, we average over three trials for all experiments and report accuracy with its standard deviation. We perform utility and safety evaluations via API queries, which do not require local GPU acceleration. For the LLMmap method, we utilize an NVIDIA DGX A100 with GPU acceleration to train the model fingerprint database. Detailed model configurations and API parameters are provided in Section A.1.

4.2 Utility Evaluation

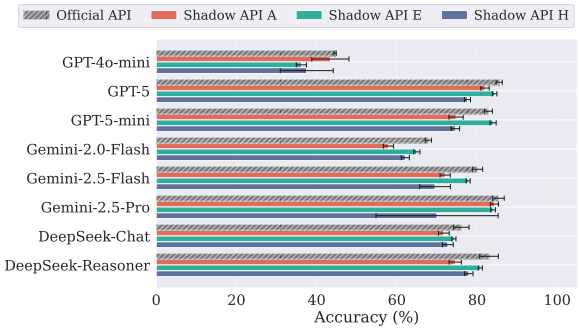
Benchmarks and Methodology. We assess model utility across scientific and sensitive domains (Gemini, 2025; DeepSeek-AI, 2025; Qwen, 2025). For the science domain, we employ the AIME 2025 benchmark (of Problem Solving, 2025), which tests competition-level mathematics, and the GPQA (Diamond) benchmark (Rein et al., 2023), targeting PhD-level scientific questions. For the sensitive domain, we focus on high-risk fields, i.e., medical and legal fields. In the medical field, following prior work (Tang et al., 2025), we use the MedQA (USMLE) dataset (Jin et al., 2021), covering diagnosis, treatment, and medical concepts. In the legal field, we select LegalBench (Guha et al., 2023), specifically the Scalr subset Rule-Application and Rule-Conclusion involving reasoning, consistent with Chu et al. (2025); Hu et al. (2025).

Implementation Details. To ensure fair comparison, we adopt EvalScope prompt templates (Contributors, 2025) for the AIME 2025 and GPQA benchmarks; for the MedQA (USMLE) benchmark, we use the Hulu-Med multiple-choice template (Jiang et al., 2025a), and the original task-specific prompts for LegalBench (Scalr) (Guha et al., 2023). We obtain all results using the model configurations detailed in Table 3.

Science Domain Performance. The official API generally establishes the performance upper bound, as illustrated in Figure 3. Among the shadow APIs, shadow API E exhibits exceptional consistency, maintaining a minimal average divergence 2.64% from the official. In some instances, such as with GPT-5-mini on GPQA, it even marginally outperforms the official API by 1.18%. In contrast, official APIs exhibit minimal performance variance, whereas shadow APIs show substantially higher variability. In particular, shadow APIs A and H display pronounced divergences from official APIs, with average accuracy gaps of 9.81% and 6.46%, respectively. Although these shadow APIs perform comparably on non-reasoning tasks, their performance degrades markedly on reasoning-oriented models. Notably, on the AIME 2025 benchmark, shadow API A suffers severe accuracy drops, with deficits 40.00% for Gemini-2.5-Pro and 38.89% for DeepSeek-Reasoner. These results indicate that advanced reasoning capabilities are significantly compromised in shadow APIs A and H, resulting in substantial deviations from their claimed parity with official API performance.



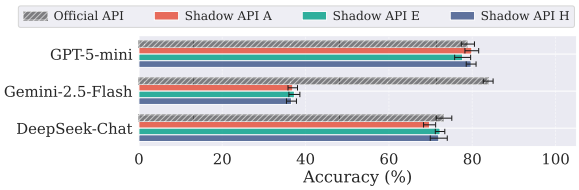
(a) AIME 2025



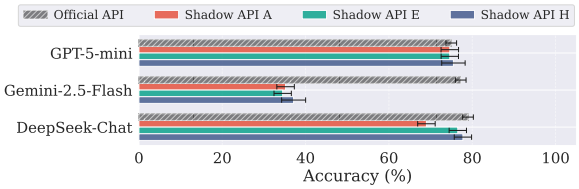
(b) GPQA (Diamond)

Figure 3: Performance comparison on (a) AIME 2025 and (b) GPQA benchmarks across official and shadow APIs.

Sensitive Domain Performance. For sensitive domain, including medicine field (MedQA (USMLE)) and law field (LegalBench (Scalr)), performance results are shown in Figure 4a and Figure 4b. Shadow APIs A, E, and H exhibit average accuracy drops of



(a) MedQA (USMLE)



(b) LegalBench (Scalr)

Figure 4: Performance comparison on (a) MedQA (USMLE) and (b) LegalBench (Scalr) benchmarks across official and shadow APIs.

16.96%, 15.71%, and 14.75%, respectively. Performance on Gemini-2.5-Flash degrades across all shadow APIs. In particular, its accuracy on MedQA decreases sharply from an official score of 83.82% to an average of 36.95%, corresponding to a substantial performance deficit of 46.51%–47.21%. A consistent collapse is observed on LegalBench, where all shadow APIs lag behind the official endpoints by 40.10% to 42.73%. DeepSeek-Chat also demonstrates instability, particularly in the legal field. While shadow APIs E and H remain relatively stable, shadow API A exhibits a significant accuracy drop of 9.98% on LegalBench.

Table 1 illustrates critical failure instances in high-stakes domains: shadow APIs confuse HIV diagnostic protocols in the medical field and misinterpret juror honesty precedents in the legal field. As quantified in Table 5 and Table 6 (see Appendix B), this unreliability is statistically significant across both fields. For example, shadow APIs of Gemini-2.5-Flash fail to reproduce the official API's correct answers in nearly half of the evaluated cases, indicating that reliance on shadow APIs for professional guidance poses severe safety risks.

4.3 Safety Evaluation

Datasets. We consider two widely used benchmarks, JailbreakBench (Chao et al., 2024) and AdvBench (Zou et al., 2023), which respectively contain 520 and 100 harmful requests covering different categories (such as deception, discrimination, and physical harm) to evaluate the safety capabilities of LLMs in the face of unsafe user requests. For AdvBench, we use the subset provided by Chao et al. (2023).

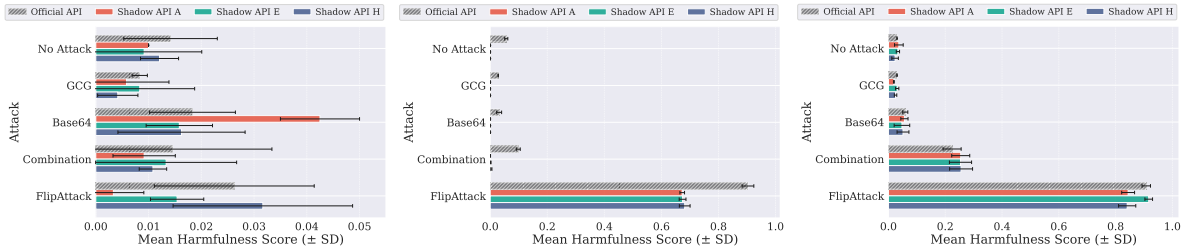
Table 1: Medical and legal fields failure response examples.

Benchmark	Question	Official API	Shadow API
MedQA (USMLE)	Perinatal HIV screening during labor... identifying confirmatory test.	HIV-1/HIV-2 antibody differentiation immunoassay ✓	**A, E, H:** Determines the genotype of the virus ✗
LegalBench (Scalr)	Whether Rule 606(b) permits juror testimony about deliberations to prove dishonesty during voir dire.	Rule 606(b) applies/bars testimony even for voir dire lies ✓	**A, H:** Confuses "new trial" standard with admissibility ✗ **E:** Cites unrelated habeas rule ✗

Implementation Details. We employ four distinct jailbreak attacks to evaluate the safety: GCG (Zou et al., 2023), Base64 (Wei et al., 2023), Combination (Wei et al., 2023), and FlipAttack (Liu et al., 2024). For GCG, we use LLaMA3-8B to generate a universal suffix and then transfer it to other LLMs. For Base64 and Combination, we follow the settings for Base64 and `combination_1` in Wei et al. (2023). For FlipAttack, we use its well-performed "flip char in sentence" mode.

Metrics. Following prior work such as PAIR (Chao et al., 2023), TAP (Mehrotra et al., 2024), and JAIL-CON (Jiang et al., 2025b), we introduce a lightweight judge model to output the *harmfulness score* and verify whether the generated answer is valid and harmful. Specifically, based on Souly et al. (2024), we build the evaluator with GPT-4o-mini and the rubric-based prompt template from StrongREJECT. A higher harmfulness score indicates a more harmful answer, i.e., lower safety.

Experimental Results. As shown in Figure 5, we observe a concerning, inconsistent behavior on JailbreakBench, where shadow APIs deviate unpredictably from official baselines. Specifically, for GPT-5-Mini under the Base64 attack (Figure 5a), shadow API A yields a harmfulness score of 0.04, which is double the official API's score of 0.02. Similarly, inconsistencies persist in the FlipAttack, where shadow API A and E significantly underestimate the risk. For Gemini-2.5-Flash (Figure 5b), the results show an underestimation of risk by all shadow APIs, which are safer than the official API across all attacks, particularly against FlipAttack (the official API reaches a high harmfulness score of 0.90, all shadow APIs around 0.67 and 0.68, resulting in a significant gap of approximately 0.23.) In the case of DeepSeek-Chat (Figure 5c), the shadow APIs exhibit smaller differences compared to GPT-5-Mini and Gemini-2.5-Flash, but differences from the official API still exist. For



(a) GPT-5-Mini (b) Gemini-2.5-Flash (c) DeepSeek-Chat

Figure 5: Safety performance comparison on the JailbreakBench dataset.

Table 2: Fingerprinting results via LLMmap matched model and mean cosine distance D with standard. Colors denote ratio vs official, Green ($D \leq 1.2\times$ baseline), Yellow ($D > 1.2\times$ baseline), and Red (incorrect model).

Model Family	Model	Official API *(Baseline)*	Shadow API A *(Standard)*	Shadow API E *(Standard)*	Shadow API H *(Standard)*
GPT	GPT-4o-mini	**gpt-4o-mini-0718** $[D: 20.01_{\pm 0.96}]$	gpt-4o-2024-05-13 $[D: 16.66_{\pm 2.40}]$	gpt-4o-mini-0718 $[D: 19.18_{\pm 1.41}]$	Qwen2.5-7B $[D: 17.43_{\pm 2.07}]$
	GPT-5	**gpt-5-2025-08-07** $[D: 13.14_{\pm 0.34}]$	glm-4-9b-chat $[D: 20.88_{\pm 2.81}]$	glm-4-9b-chat $[D: 23.50_{\pm 0.83}]$	gpt-5-2025-08-07 $[D: 16.24_{\pm 3.84}]$
	GPT-5-mini	**gpt-5-mini-2025-08-07** $[D: 14.57_{\pm 3.82}]$	gpt-5-mini-2025-08-07 $[D: 18.63_{\pm 2.72}]$	gpt-5-mini-2025-08-07 $[D: 16.01_{\pm 3.59}]$	gpt-5-2025-08-07 $[D: 18.04_{\pm 1.76}]$
Gemini	Gemini-2.0-Flash	**gemini-2.0-flash** $[D: 17.54_{\pm 2.28}]$	gemini-2.5-flash $[D: 17.02_{\pm 2.64}]$	gemini-2.5-flash $[D: 14.10_{\pm 1.28}]$	gemini-2.0-flash $[D: 18.49_{\pm 2.50}]$
	Gemini-2.5-Flash	**gemini-2.5-flash** $[D: 15.19_{\pm 2.71}]$	gemini-2.5-flash $[D: 19.78_{\pm 1.92}]$	gemini-2.5-flash $[D: 15.41_{\pm 3.37}]$	gemini-2.5-flash $[D: 17.51_{\pm 2.22}]$
	Gemini-2.5-Pro	**gemini-2.5-pro** $[D: 18.04_{\pm 0.93}]$	gemini-2.5-pro $[D: 18.04_{\pm 0.61}]$	gemini-2.5-pro $[D: 17.37_{\pm 1.76}]$	gemini-2.5-pro $[D: 17.66_{\pm 1.56}]$
DeepSeek	DeepSeek-Chat	**deepseek-chat** $[D: 11.83_{\pm 0.79}]$	deepseek-v3-0324 $[D: 19.77_{\pm 3.78}]$	deepseek-chat $[D: 21.28_{\pm 1.57}]$	gemma-2-9b-it $[D: 23.40_{\pm 1.05}]$
	DeepSeek-Reasoner	**deepseek-reasoner** $[D: 17.04_{\pm 3.25}]$	deepseek-chat $[D: 19.02_{\pm 1.53}]$	deepseek-reasoner $[D: 21.19_{\pm 0.34}]$	deepseek-chat $[D: 22.68_{\pm 2.69}]$

instance, shadow APIs A and H generate more (less) harmful content under the attack Combination (FlipAttack). We also have similar results on the AdvBench, with detailed experimental results provided in Appendix C. These findings imply that relying on shadow APIs for safety evaluation is unreliable, as they may fail to reproduce the behaviors of the official endpoints.

RQ2 Take-Aways: Performance evaluation shows shadow APIs could not be replacements for official ones, as they often break on reasoning, and they are unreliable in medical/legal tasks and safety evaluation.

5 Model Verification

To address **RQ3**, we move beyond indirect performance signals and directly verify the identity of the served models using model fingerprinting techniques and metadata comparison.

5.1 Fingerprinting-Based Detection

Methodology. To identify the specific LLMs operating behind shadow API service providers, we employ LLMmap (Pasquini et al., 2025), an active fingerprinting framework. This framework classifies models by analyzing responses to a curated set of queries and computing the cosine distance between model outputs and a reference database. We conduct our fingerprinting experiments following the configuration in Pasquini et al. (2025), using the default query strategy and extending the framework with the new LLM list. For GPT-5-mini, GPT-5, Gemini-2.5-Flash, and Gemini-2.5-Pro, we remove unsupported parameters and configure each model to use its default medium reasoning effort. The full list of LLMs used is provided in Appendix E.

Results. We observe significant variance in the identity reliability of shadow APIs across the 24 evaluated endpoints, as summarized in Table 2. Specifically, 45.83% of the endpoints fail fingerprint verification, and an additional 12.50% exhibit substantial cosine distance deviations from the corresponding official models. At the family level, the GPT and DeepSeek families show frequent identity mismatches. Even when a shadow API is correctly identified within the same model family, it often exhibits inflated cosine distances (e.g., GPT-5-mini in shadow API A, $D = 18.63_{\pm 2.72}$ versus an official baseline of $14.57_{\pm 3.82}$). By contrast, the Gemini family exhibits relatively high stability for Gemini-2.5-Pro, which maintains consistent cosine

distances ($D \approx 17.37$–18.04) across all providers.

Shadow providers frequently employ two primary forms of deception, as supported by the following evidence. First, premium proprietary models exhibit response patterns that more closely align with cheaper open-source alternatives. For instance, the behavior of GPT-4o-mini in shadow API H deviates toward Qwen2.5-7B, while GPT-5 in APIs A and E shows fingerprint signatures resembling GLM-4 or DeepSeek-V3. Second, specialized or reasoning models may be served by non-reasoning models. Requests for the thinking-mode DeepSeek-Reasoner through APIs A and H instead behave like the non-thinking DeepSeek-Chat. Similarly, Gemini-2.0-Flash is often misidentified as Gemini-2.5-Flash. Taken together, this evidence points to identity inconsistencies that are difficult for end users to reliably verify.

5.2 Meta Information Analysis

We analyze inference latency time and token counts to identify inconsistent behaviors in Appendix D. Official APIs typically exhibit consistent inference latency and token counts for the same question, while shadow APIs exhibit irregular spikes. Standard deviation analysis corroborates this instability, revealing that shadow APIs frequently exhibit volatility exceeding even $2.0\times$ the official ones.

RQ3 Take-Aways: Model verification provides direct evidence of deception that nearly half of the shadow APIs could not pass fingerprint verification, and their inference latency and token counts deviate from official ones.

6 Joint Analysis

Based on the combined results from performance evaluation (Section 4) and model verification (Section 5), we analyze how model identity relates to observed performance divergence. In some cases, matching model identity coincides with consistent behavior. For example, when model identity matches and behavior remains stable, shadow APIs can perform closely to official endpoints, as observed for GPT-5-mini in shadow API E, where fingerprinting matches the claimed model and performance in sensitive domains remains largely unchanged. Conversely, when model identity does not match, behavior often degrades accordingly. During reasoning evaluation, identity mismatch is strongly associated with reasoning collapse, as when DeepSeek-Reasoner served by shadow API A fingerprints as DeepSeek-Chat and its AIME 2025 accuracy drops significantly, with a similar pattern observed for GPT-4o-mini in shadow API H. However, this consistency is not stable across shadow APIs. Model substitution does not always manifest as immediate performance degradation. GPT-5 in shadow APIs A and E fingerprints as glm-4-9b-chat, yet its AIME accuracy declines moderately, suggesting that substitution can be harder to detect in some scientific benchmarks and may become more visible under higher reasoning pressure or in tasks with strict correctness constraints. Likewise, matching model identity does not guarantee faithful behavior. Gemini-2.5-Flash illustrates this, across shadow APIs A, E, and H, fingerprinting matches the claimed model family with cosine distances close to the official APIs, yet accuracy in sensitive domains drops sharply, indicating that identity checks alone cannot ensure behavioral consistency.

7 Suggestion

To safeguard the future of LLM evaluation, we propose the following actionable suggestions. Given the high volatility of shadow services, authors must explicitly disclose the API endpoint used, record specific model versions, the service provider, and the date of access. Researchers can no longer treat third-party APIs as trusted black boxes, and the community should adopt active auditing protocols. Utilize active fingerprinting tools like LLMmap to verify that the served model's output distribution matches official baselines. For specialized tasks, researchers should verify the reasoning results. Conference organizers should update reviewer guidelines to flag usage of unverified third-party endpoints as a risk to the paper's validity.

8 Conclusion

In this work, we presented the first systematic audit of the shadow API, a widely used but unverified access for frontier LLMs. Through a comprehensive evaluation across providers, benchmarks, and multi-dimensional benchmarks, behavioral fingerprinting, we demonstrated that shadow APIs exhibit significant performance divergence and identity inconsistency. Relying on shadow APIs for evaluation is unreliable, as they may fail to reproduce the behaviors of the official endpoints. Taken together, these findings reveal deceptive model claims in shadow APIs and demonstrate that they cannot be treated as reliable official models.

Limitations

Despite providing a systematic audit of the shadow APIs, our study is subject to several limitations.

The shadow API market is characterized by extreme opacity. Providers frequently switch upstream model sources, alter routing strategies, or cease operations without notice. Our findings reflect a snapshot of the shadow API within a specific time window (September to December 2025). As official countermeasures intensify and market dynamics evolve, the behavioral characteristics of these services may drift significantly in the future.

In the absence of ground truth regarding the backend infrastructure of shadow APIs, our investigation relies on performance metrics, meta-information, and established auditing frameworks such as LLMmap to detect model substitution and performance inconsistencies. While we can detect model identity mismatches with high confidence, we cannot definitively distinguish between certain fine-grained implementation details.

Although we collect the 17 providers based on academic citations and GitHub stars, this may not represent the full shadow API market. More shadow API services may exist, which may employ technical architectures or deceptive strategies distinct from those observed in this study. Also, our work could not cover all LLM families, so we mainly focus on three representative model families (eight LLMs). We believe our audit pipeline and findings have the potential to be extrapolated to more LLM families and look forward to future work focusing on wider model families.

Moreover, while we specify exact version snapshots for official models wherever possible, the official APIs of proprietary models may undergo undisclosed fine-tuning. This implies that the official baselines themselves may exhibit minor temporal variance, potentially posing challenges for precise reproducibility over extended periods. To mitigate the potential impact of temporal drift and inherent stochasticity in official APIs, we conducted multiple independent runs for each experiment and reported the variance, ensuring that our comparative results remain robust against fluctuations in the baseline models.

Ethical Considerations

In this work, we adhere to responsible AI research guidelines. Our study involved purchasing and querying unofficial shadow API services, which may violate the terms of service of official providers (e.g., OpenAI, Google, DeepSeek). These interactions were conducted solely for the purpose of auditing and transparency. We do not endorse, promote, or encourage the use of these unauthorized services. To minimize potential negative impacts on official infrastructure, we strictly limited our query volume to the minimum necessary for statistical validation of model identity and performance. Furthermore, all requests to official APIs were conducted exclusively from authorized geographic regions to ensure full compliance with the providers' regional access policies.

To avoid serving as an advertisement for illicit services and to mitigate legal risks for individual operators, we have applied strict anonymization to all audited shadow APIs. Our goal is to expose systemic risks within the shadow API market rather than to target specific individuals. Furthermore, regarding our prevalence analysis, we have strictly de-identified all specific academic papers, authors, and institutions found to be using shadow APIs. Our objective is to highlight systemic reproducibility within the community.

Our safety evaluation involved the use of jailbreak attacks to test model guardrails. To prevent misuse, we follow best practices in safety research: we do not release specific examples of successfully generated harmful content or the exact adversarial prompts used to bypass safety filters. We report only aggregated metrics, harmfulness score, and store all sensitive outputs on secure, access-controlled local servers.

To mitigate potential harm and foster a healthier market, we have adhered to responsible disclosure principles. We are committed to responsibly reporting our findings to official model providers as well as the authors of papers that use shadow APIs. We have received partial acknowledgment from them.

References

AAAI. 2025. AAAI-26 Review Process Update: Scale, Integrity Measures, and Pathways to Sustainability. https://aaai.org/conference/aaai/aaai-26/review-process-update/. Accessed: 2025-12-06.

ACL. 2024. ACL 2024 Main Conference Accepted Papers. https://2024.aclweb.org/program/main_conference_papers/. Accessed: 2025-12-06.

DeepSeek AI. 2025. DeepSeek API Documentation: Parameter Settings. https://github.com/QuantumNous/new-api. Accessed: 2025-12-06.

Anthropic. 2025. Updating Restrictions of Sales to Unsupported Regions. https://tinyurl.com/4xawc72e. Accessed: 2025-12-06.

Archana. 2025. The Unsustainable Economics of LLM APIs. https://tinyml.substack.com/p/the-unsustainable-economics-of-llm. Accessed: 2025-12-06.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. *CoRR abs/2310.05130*.

Bloomberg. 2024. OpenAI Taking Steps to Block China's Access to Its AI Tools. https://tinyurl.com/yb2xcez9. Accessed: 2025-12-06.

Hudson Borges, Andre Hora, and Marco Tulio Valente. 2016. Understanding the Factors that Impact the Popularity of GitHub Repositories. In *International Conference on Software Maintenance and Evolution (ICSME)*. IEEE.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR abs/2303.12712*.

Will Cai, Tianneng Shi, Xuandong Zhao, and Dawn Song. 2025. Are You Getting What You Pay For? Auditing Model Substitution in LLM APIs. *CoRR abs/2504.04715*.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 55005–55029. NeurIPS.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking Black Box Large Language Models in Twenty Queries. *CoRR abs/2310.08419*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *CoRR abs/2305.05176*.

Xu Chu, Zhijie Tan, Hanlin Xue, Guanyu Wang, Tong Mo, and Weiping Li. 2025. Domaino1s: Guiding LLM Reasoning for Explainable Answers in High-Stakes Domains. *CoRR abs/2501.14431*.

EvalScope Contributors. 2025. EvalScope: Supported Datasets for LLM Evaluation. https://evalscope.readthedocs.io/en/v1.0.0/get_started/supported_dataset/llm.html. Accessed: 2025-12-06.

Cyberspace Administration of China. 2023. Interim Measures for the Management of Generative Artificial Intelligence Services. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm. Accessed: 2025-12-06.

DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *CoRR abs/2512.02556*.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *CoRR abs/2502.05151*.

Gao, Irena, Liang, Percy, Guestrin, and Carlos. 2024. Model Equality Testing: Which Model Is This API Serving? *CoRR abs/2410.20247*.

Gemini. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *CoRR abs/2507.06261*.

Google. 2025. Available regions for Google AI Studio and Gemini API. https://ai.google.dev/gemini-api/docs/available-regions. Accessed: 2025-12-06.

Google DeepMind. 2025. Nano Banana Pro. https://deepmind.google/models/gemini-image/pro/. Accessed: 2026-01-03.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *Advances in neural information processing systems*, 36:44123–44279.

Yinghao Hu, Yaoyao Yu, Leilei Gan, Bin Wei, Kun Kuang, and Fei Wu. 2025. Evaluating Test-Time Scaling LLMs for Legal Reasoning: OpenAI o1, DeepSeek-R1, and Beyond. *CoRR abs/2503.16040*.

ICLR. 2024. International Conference on Learning Representations (ICLR) 2024 Accepted Papers. https://iclr.cc/virtual/2024/papers.html. Accessed: 2025-12-06.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, and 6 others. 2025a. Hulu-Med: A Transparent Generalist Model towards Holistic Medical Vision-Language Understanding. *CoRR abs/2510.08668*.

Yukun Jiang, Mingjie Li, Michael Backes, and Yang Zhang. 2025b. Adjacent Words, Divergent Intents: Jailbreaking Large Language Models via Task Concurrency. *CoRR abs/2510.21189*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*.

Keegan Keplinger. 2024. Hackers Are Stealing GenAI Credentials, So What Sensitive Company Data Are They Getting Their Hands On? https://tinyurl.com/bdhvry4n. Accessed: 2025-12-06.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. 2019. Thieves on Sesame Street! Model Extraction of BERT-based APIs. *CoRR abs/1910.12366*.

LatticeFlow. 2024. CVPR'24 in Numbers. https://latticeflow.ai/news/cvpr24-in-numbers. Accessed: 2025-12-06.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2022. Holistic Evaluation of Language Models. *CoRR abs/2211.09110*.

Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, and Bryan Hooi. 2024. FlipAttack: Jailbreak LLMs via Flipping. *CoRR abs/2410.02832*.

Hope McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. 2025. Your Large Language Models Are Leaving Fingerprints. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 85–95.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 61065–61105. NeurIPS.

Art of Problem Solving. 2025. AIME Problems and Solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2025-12-06.

OpenAI. 2023. GPT-4 Technical Report. *CoRR abs/2303.08774*.

OpenAI. 2024. ChatGPT and API Services in Unsupported Countries and Territories. https://tinyurl.com/69nw2cn3. Accessed: 2025-12-06.

OpenAI. 2025a. OpenAI Services Agreement. https://openai.com/policies/services-agreement. Accessed: 2025-12-06.

OpenAI. 2025b. Reasoning Models - OpenAI API Documentation. https://platform.openai.com/docs/guides/reasoning. Accessed: 2025-12-06.

OpenAI. 2025c. Supported Countries and Territories. https://platform.openai.com/docs/supported-countries. Accessed: 2025-12-06.

Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. 2025. LLMmap: Fingerprinting for large language models. In *USENIX Security Symposium (USENIX Security)*, pages 299–318. USENIX.

QuantumNous. 2025. New API: An AI Model Interface Management and Distribution System Based on One API. https://github.com/QuantumNous/new-api. Accessed: 2025-12-06.

Qwen. 2025. Qwen3 Technical Report. *CoRR abs/2505.09388*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *CoRR abs/2311.12022*.

Sebők Miklós and Kiss Rebeka. 2025. Unlocking Large Language Models via API: Capabilities, Access, and Practical Considerations. https://tinyurl.com/35juu2n7. Accessed: 2025-12-06.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2023. Large Language Models Encode Clinical Knowledge. *Nature*, 620(7972):172–180.

Quanpeng Song. 2025. One API: An OpenAI-compatible All-in-one AI API Gateway. https://github.com/songquanpeng/one-api. Accessed: 2025-12-06.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for Empty Jailbreaks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 125416–125440. NeurIPS.

Sun, Yifan, Li, Yuhang, Zhang, Yue, Jin, Yuchen, Zhang, and Huan. 2024. SVIP: Towards Verifiable Inference of Open-source Large Language Models. *CoRR abs/2410.22307*.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, Arman Cohan, and Mark Gerstein. 2025. MedAgentsBench: Benchmarking Thinking Models and Agent Frameworks for Complex Medical Reasoning. *CoRR abs/2503.07459*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiqiu Xia, Lang Zhu, Bingzhe Li, Feng Chen, Qiannan Li, Chunhua Liao, Feiyi Wang, and Hang Liu. 2025. Analyzing 16,193 LLM Papers for Fun and Profits. *CoRR abs/2504.08619*.

Zhikang Xie, Weilin Wan, Peizhu Gong, Weizhong Zhang, and Cheng Jin. 2025. Advanced black-box tuning of large language models with limited api calls. *CoRR abs/2511.10210*.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. *CoRR abs/2305.17359*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A Survey of Large Language Models. *CoRR abs/2303.18223*, 1(2).

Zilan Qian. 2025. How to Use Banned US Models in China: The Grey Market for American LLMs. https://www.chinatalk.media/p/the-grey-market-for-american-llms. Accessed: 2025-12-06.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. *CoRR abs/2307.15043*.

A Configurations Details

A.1 Model Configurations

Table 3 details the specific configurations for all evaluated models across official and shadow APIs. To ensure reproducibility, we set the random seed at 42 and the temperature at 0 for all models compatible with these settings, in accordance with reasoning task guidelines (AI, 2025), while omitting these parameters for unsupported endpoints. For the GPT-5 model family and other reasoning models, we specifically monitor their support for reasoning parameters. For all reasoning models, we employ the medium reasoning effort in our subsequent analysis, according to the default configuration (OpenAI, 2025b). The table also highlights the discrepancies in which shadow APIs may enforce fixed behaviors that deviate from official baselines. The Ratio column denotes the multiplicative factor of the shadow API service price relative to the official provider for input/output tokens. Logprobs indicates whether the API can return log probability outputs. Pricing is normalized to USD per 1 million tokens for comparison and accessed on December 6, 2025.

A.2 Anonymized Details of API Service Providers

Table 4 provides a complete list of the 17 unofficial API service providers identified in our study, with their names anonymized for privacy and ethical considerations.

A.3 Shadow APIs Profiles

Shadow API A is selected as the first ranking in both categories, which is operated by an individual and does not employ open-source LLM API management or distribution systems. Shadow API E and H are selected for their stability, as they rank within the top for both citations and stars. Regarding shadow API E, the provider is also an individual and utilizes an open-source LLM API management and distribution system (NewAPI (QuantumNous, 2025)) as its deployment template. Shadow API E aggregates over 54 upstream providers, some of which are explicitly labeled as having undisclosed origins. Some of the endpoints are categorized as reverse-engineered, web-based, capability-degraded, distilled, or unstable yet cost-effective variants. Given this diversity, we specifically select endpoints explicitly labeled as originating from the official. For the shadow API H, the provider is a

Table 3: Comparison of model configurations across different APIs.

Model Family	Model Name	Channel	API Model ID	Temp.	Seed	Logprobs	Price (In/Out)	Ratio
GPT	GPT-4o-Mini	Official	gpt-4o-mini-2024-07-18	0	42	●	$0.15 / $0.60	1.00 / 1.00
		Shadow API A	gpt-4o-mini	0	42	●	$0.15 / $0.60	1.00 / 1.00
		Shadow API E	gpt-4o-mini-2024-07-18	0	42	●	$0.11 / $0.43	0.73 / 0.72
		Shadow API H	gpt-4o-mini-2024-07-18	0	42	●	$0.16 / $0.65	1.09 / 1.09
	GPT-5	Official	gpt-5-2025-08-07	N/A	42	○	$1.25 / $10.00	1.00 / 1.00
		Shadow API A	gpt-5	0	42	◐	$1.25 / $10.00	1.00 / 1.00
		Shadow API E	gpt-5-2025-08-07	0	42	○	$0.89 / $7.14	0.71 / 0.71
		Shadow API H	gpt-5-2025-08-07	0	42	○	$1.36 / $10.90	1.09 / 1.09
	GPT-5-Mini	Official	gpt-5-mini-2025-08-07	N/A	42	○	$0.25 / $2.00	1.00 / 1.00
		Shadow API A	gpt-5-mini	0	42	◐	$0.25 / $2.00	1.00 / 1.00
		Shadow API E	gpt-5-mini-2025-08-07	0	42	○	$0.18 / $1.43	0.72 / 0.72
		Shadow API H	gpt-5-mini	0	42	○	$0.27 / $2.18	1.09 / 1.09
Gemini	Gemini-2.0-Flash	Official	gemini-2.0-flash	0	N/A	●	$0.10 / $0.40	1.00 / 1.00
		Shadow API A	gemini-2.0-flash	0	42	◐	$0.71 / $2.90	7.10 / 7.25
		Shadow API E	gemini-2.0-flash	0	42	◐	$0.07 / $0.29	0.71 / 0.73
		Shadow API H	gemini-2.0-flash	0	42	◐	$0.11 / $0.44	1.09 / 1.09
	Gemini-2.5-Flash	Official	gemini-2.5-flash	0	N/A	●	$0.30 / $2.50	1.00 / 1.00
		Shadow API A	gemini-2.5-flash	0	42	◐	$0.09 / $2.00	0.29 / 0.80
		Shadow API E	gemini-2.5-flash	0	42	◐	$0.21 / $1.79	0.70 / 0.72
		Shadow API H	gemini-2.5-flash	0	42	◐	$0.33 / $2.73	1.09 / 1.09
	Gemini-2.5-Pro	Official	gemini-2.5-pro	0	N/A	●	$1.25 / $10.00	1.00 / 1.00
		Shadow API A	gemini-2.5-pro	0	42	◐	$1.00 / $5.70	0.80 / 0.57
		Shadow API E	gemini-2.5-pro	0	42	◐	$0.89 / $7.14	0.71 / 0.71
		Shadow API H	gemini-2.5-pro	0	42	◐	$1.36 / 10.90	1.09 / 1.09
DeepSeek	DeepSeek-Chat	Official	deepseek-chat	0	42	●	$0.28 / $0.42	1.00 / 1.00
		Shadow API A	deepseek-v3.2	0	42	◐	$0.17 / $0.26	0.61 / 0.62
		Shadow API E	DeepSeek-V3.2-nothinking	0	42	●	$1.43 / $2.14	5.11 / 5.10
		Shadow API H	deepseek-chat	0	42	◐	$0.22 / $0.34	0.80 / 0.80
	DeepSeek-Reasoner	Official	deepseek-reasoner	0	42	●	$0.28 / $0.42	1.00 / 1.00
		Shadow API A	deepseek-v3.2-thinking	0	42	●	$0.17 / $0.26	0.61 / 0.62
		Shadow API E	DeepSeek-V3.2-thinking	0	42	●	$1.43 / $2.14	5.11 / 5.10
		Shadow API H	deepseek-reasoner	0	42	●	$0.22 / $0.34	0.80 / 0.80

●: Accept logprobs with output; ◐: Accept logprobs but no output; ○: Not accept logprobs.
Ratio: The Ratio column denotes the multiplicative factor of the shadow API service price relative to the official provider for input/output tokens. Ratio < 1 (> 1) indicates the shadow API is cheaper (more expensive) than the official ones.

Table 4: Shadow API service providers with anonymized names.

Shadow API	Anonymized Name
A	C****E
B	Y****I
C	X****I
D	G****S
E	Q****O
F	O****B
G	D****I
H	Z****G
I	C****I
J	O****D
K	V****I
L	A****S
M	B****I
N	A****X
O	A****S
P	A****9
Q	A****I

corporate entity (Internet Content Provider filing) rather than an individual operator, which does not rely on an open-source LLM API management and distribution system. Notably, the shadow API B is deployed by a laboratory at a prestigious Chinese university, built on NewAPI, where part of the models listed in its internal large language Model Marketplace are explicitly marked with Unknown sources.

A.4 OpenRouter Ranking

Figure 6 presents the detailed snapshot of the OpenRouter, November 2025. This ranking highlights the most widely used models by token consumption.

B Detailed Discrepancy

Table 5 and Table 6 present a detailed discrepancy analysis between official and shadow APIs, categorizing their response across MedQA (USMLE) and LegalBench (Scalr). We report results based on the first execution.

Particularly, Gemini-2.5-Flash exhibits performance collapse across all shadow APIs, with an average consistency rate of $\sim$ 51.48% in medi-

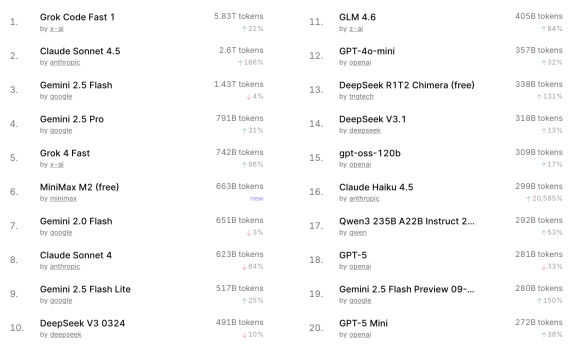


Figure 6: LLM token usage rankings on OpenRouter.

cal and legal fields. This overwhelmingly high count of official-correct-only instances confirms that shadow APIs fail to reproduce the reasoning capabilities of the official API. On DeepSeek-Chat, the advantages of this official model are reduced, but still exist as the shadow API A shows a significant degradation in the legal field (80.56% consistency), diverting sharply from the official baseline compared to other shadow providers.

C Safety Evaluation on AdvBench

We also have similar results on the AdvBench as JailbreakBench, in Figure 7, where shadow APIs deviate unpredictably from official endpoints. Specifically, for GPT-5-Mini under the Base64 attack (Figure 7a), shadow API A yields a harmfulness score of 0.05, which is $5\times$ the official API's score of 0.01. Similarly, shadow APIs A, E, and H significantly underestimate the risk in the Combination attack. For instance, Gemini-2.5-Flash (Figure 7b), the results show an underestimation of risk by all shadow APIs, which are safer than the official API across all attacks, particularly against FlipAttack (the official API reaches a high harmfulness score of 0.95, all shadow APIs around 0.82). In the case of DeepSeek-Chat (Figure 7c), the shadow APIs exhibit smaller differences compared to GPT-5-Mini and Gemini-2.5-Flash, but differences from the official API still exist. For example, shadow API H generates more harmful content than the official API under the Combination (FlipAttack) attack. In addition, the harmfulness score for shadow API A and E is comparable to the official endpoints in Base64 and Combination, but lower than the official scores in FlipAttack.

D Inference Latency Time and Token Counts

Figure 8 and Figure 9 illustrate the inference latency time and token counts on the AIME 2025. Similarly, Figure 10 and Figure 11 present the comparative analysis on the GPQA. As shown, shadow API H exhibits inference latency time and token counts that are sometimes higher and sometimes lower than those of the official API on question with GPT-4o-Mini, indicating inconsistent performance characteristics across runs.

In addition, we analyze the stability of shadow APIs compared to the official API by examining the standard deviation (SD) of inference latency time (Table 7) and token counts (Table 8). Regarding inference latency time, shadow APIs exhibit marked instability, with volatility frequently exceeding the official baseline by over $1.2\times$ or even $2.0\times$. This unpredictability is most severe in the Gemini and DeepSeek model families. In terms of token counts, GPT series models on the AIME show abnormally low variance ($< 0.8\times$), while Gemini and DeepSeek models frequently demonstrate excessive variance ($> 1.2\times$) on GPQA.

E LLMs Used for Training LLMmap

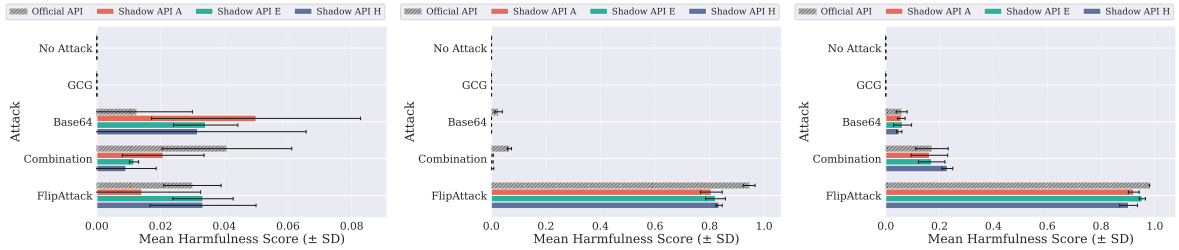
Table 9 provides a comprehensive inventory of the LLMs utilized in this study, distinguishing between the original baseline models and newly integrated trained models. All officially trained models are queried directly through the corresponding official APIs.

Table 5: Performance discrepancy between official and shadow APIs on MedQA (USMLE).

Model	Shadow API	Total	Both Correct	Both Incorrect	Official Correct Only	Shadow Correct Only	Consistency (%)
Gemini-2.5-Flash	Shadow API A	1273	448	184	619	22	49.65
	Shadow API E	1273	450	181	617	25	49.57
	Shadow API H	1273	446	186	621	20	49.65
GPT-5-Mini	Shadow API A	1273	981	232	24	36	95.29
	Shadow API E	1273	949	228	56	40	92.46
	Shadow API H	1273	977	230	28	38	94.82
DeepSeek-Chat	Shadow API A	1273	827	280	105	61	86.96
	Shadow API E	1273	905	326	27	15	96.70
	Shadow API H	1273	888	313	44	28	94.34

Table 6: Performance discrepancy between official and shadow APIs on Legalbench (Scalr).

Model	Shadow API	Total	Both Correct	Both Incorrect	Official Correct Only	Shadow Correct Only	Consistency (%)
Gemini-2.5-Flash	Shadow API A	571	187	116	254	14	53.06
	Shadow API E	571	183	116	258	14	52.36
	Shadow API H	571	197	115	244	15	54.64
GPT-5-Mini	Shadow API A	571	416	133	12	10	96.15
	Shadow API E	571	411	128	17	15	94.40
	Shadow API H	571	416	128	12	15	95.27
DeepSeek-Chat	Shadow API A	571	367	93	84	27	80.56
	Shadow API E	571	417	100	34	20	90.54
	Shadow API H	571	427	103	24	17	92.82



(a) GPT-5-Mini (b) Gemini-2.5-Flash (c) DeepSeek-Chat

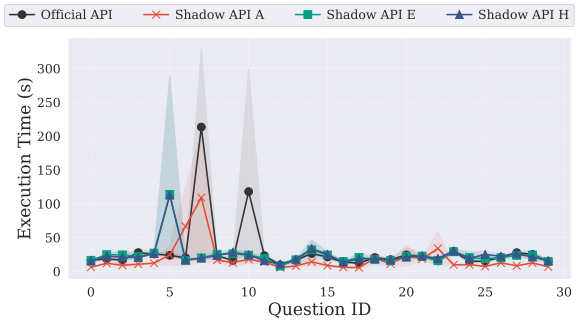
Figure 7: Safety performance comparison on the AdvBench dataset.

Table 7: Standard deviation of inference latency time (s). Colors denote ratio vs official API, Blue ($< 0.8\times$), Green ($0.8\times$ to $1.2\times$), Red ($> 1.2\times$).

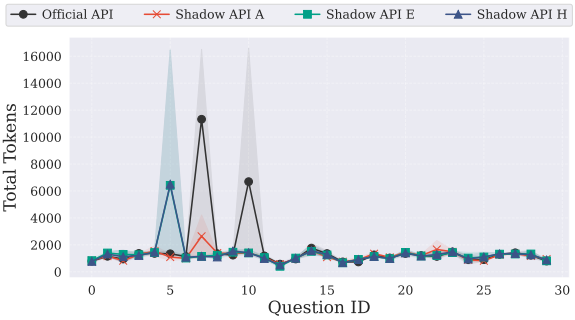
Model Family	Model Name	Benchmark	Official API	Shadow API A	Shadow API E	Shadow API H
GPT	GPT-4o-Mini	AIME	42.71	30.42	27.96	28.12
		GPQA	2.19	3.36	2.63	3.00
	GPT-5	AIME	189.04	68.62	229.20	193.95
		GPQA	32.39	45.59	40.90	36.68
	GPT-5-Mini	AIME	46.75	89.06	29.56	38.58
		GPQA	12.18	19.11	15.39	15.76
Gemini	Gemini-2.0-Flash	AIME	6.06	5.46	6.35	5.86
		GPQA	1.76	5.56	3.69	6.76
	Gemini-2.5-Flash	AIME	29.41	86.24	22.70	51.03
		GPQA	21.30	57.95	23.86	53.29
	Gemini-2.5-Pro	AIME	48.88	122.61	194.45	231.95
		GPQA	19.11	36.02	22.83	49.88
DeepSeek	DeepSeek-Chat	AIME	9.18	104.90	12.75	29.22
		GPQA	10.19	31.65	12.00	11.60
	DeepSeek-Reasoner	AIME	100.56	303.38	115.82	218.50
		GPQA	77.73	165.78	113.34	239.67

Table 8: Standard deviation of token counts. Colors denote ratio vs official API, Blue ($< 0.8\times$), Green ($0.8\times$ to $1.2\times$), Red ($> 1.2\times$).

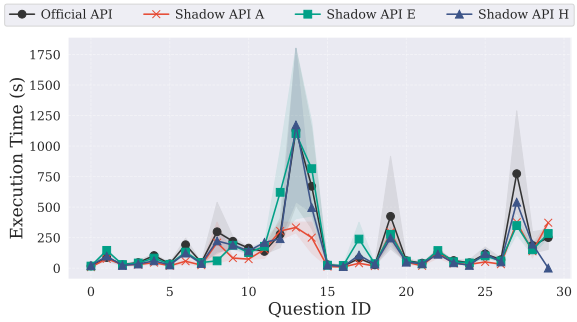
Model Family	Model Name	Benchmark	Official API	Shadow API A	Shadow API E	Shadow API H
GPT	GPT-4o-Mini	AIME	2273.87	472.75	1598.74	1585.51
		GPQA	66.13	83.53	57.68	132.32
	GPT-5	AIME	4065.04	1311.57	3919.61	2152.53
		GPQA	1598.09	1457.36	1396.30	1594.97
	GPT-5-Mini	AIME	3458.01	2026.09	1792.22	2319.84
		GPQA	892.41	1064.52	883.13	838.42
Gemini	Gemini-2.0-Flash	AIME	1243.80	1242.84	1311.17	1231.42
		GPQA	384.25	1090.25	749.74	969.43
	Gemini-2.5-Flash	AIME	8174.81	16489.45	6124.36	14164.34
		GPQA	5856.16	7029.59	6595.46	14464.05
	Gemini-2.5-Pro	AIME	3054.32	1161.70	3768.74	2905.47
		GPQA	2637.50	7430.27	3556.55	6878.14
DeepSeek	DeepSeek-Chat	AIME	307.92	492.85	383.21	1107.18
		GPQA	305.67	338.70	360.34	354.99
	DeepSeek-Reasoner	AIME	3036.03	9647.40	3544.03	3404.37
		GPQA	2314.59	6153.61	3002.75	2409.62



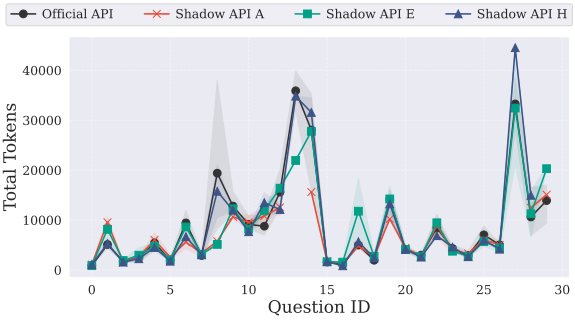
(a) AIME: GPT-4o-Mini (Time)



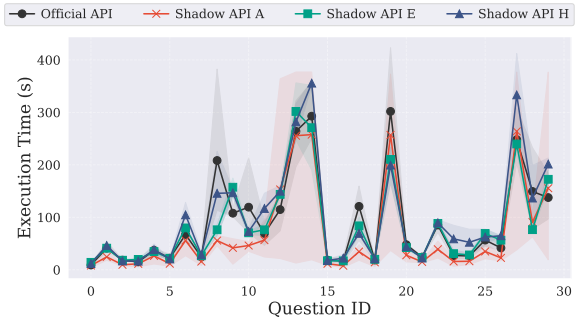
(b) AIME: GPT-4o-Mini (Token)



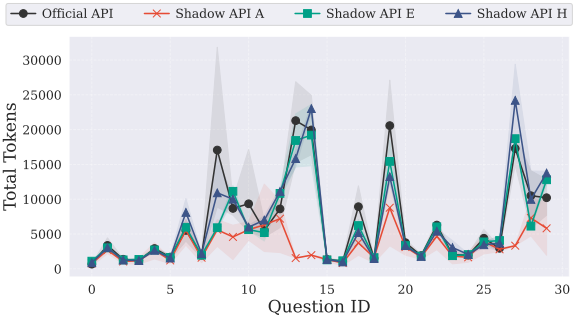
(c) AIME: GPT-5 (Time)



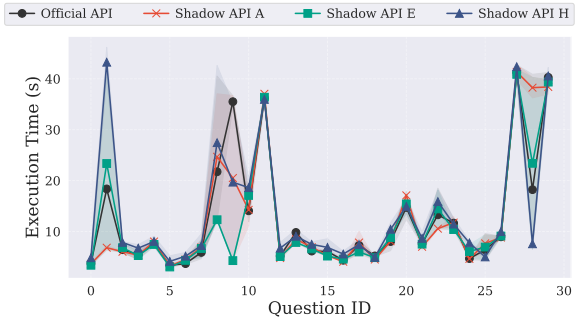
(d) AIME: GPT-5 (Token)



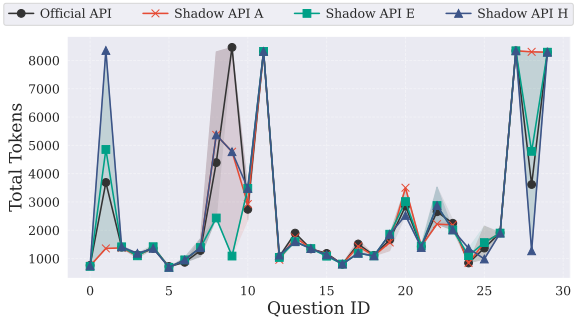
(e) AIME: GPT-5-Mini (Time)



(f) AIME: GPT-5-Mini (Token)

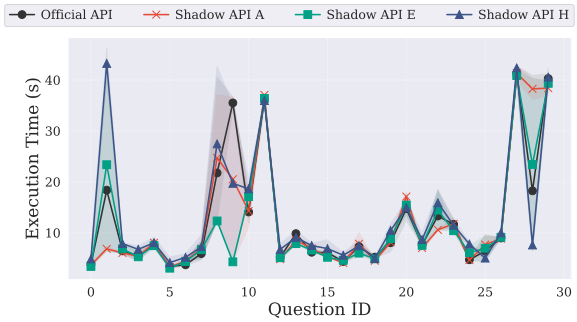


(g) AIME: Gemini-2.0-Flash (Time)

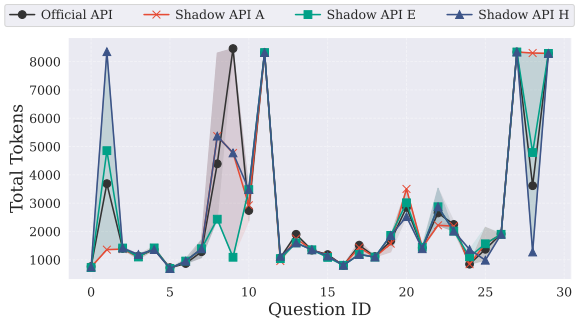


(h) AIME: Gemini-2.0-Flash (Token)

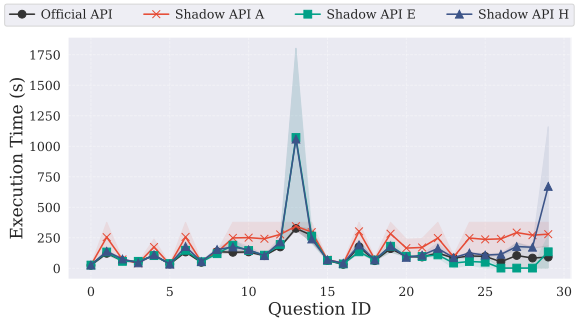
Figure 8: Comparison of inference latency time and token counts on the AIME (Part 1). Solid lines represent mean values, and shaded regions denote the range between the minimum and maximum values across three trials.



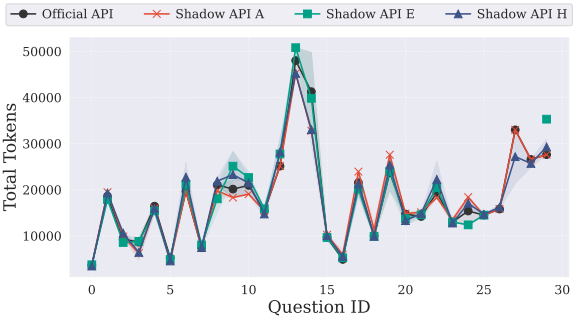
(a) AIME: Gemini-2.5-Flash (Time)



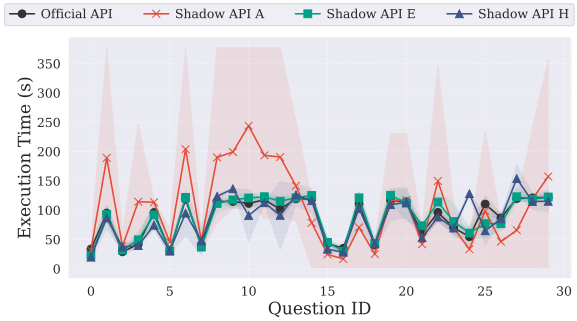
(b) AIME: Gemini-2.5-Flash (Token)



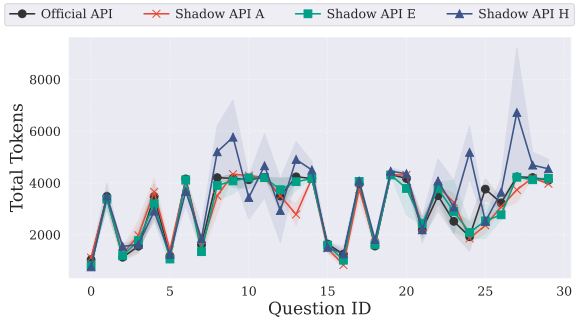
(c) AIME: Gemini-2.5-Pro (Time)



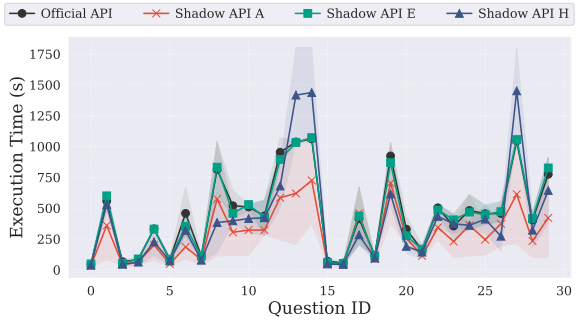
(d) AIME: Gemini-2.5-Pro (Token)



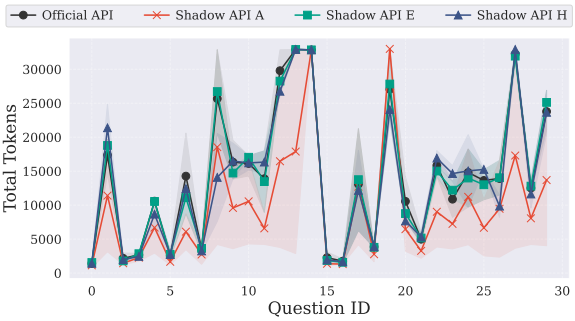
(e) AIME: DeepSeek-Chat (Time)



(f) AIME: DeepSeek-Chat (Token)

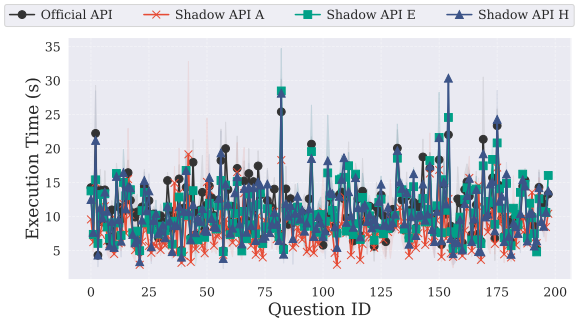


(g) AIME: DeepSeek-Reasoner (Time)

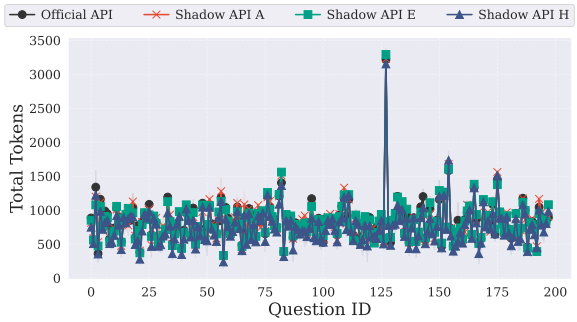


(h) AIME: DeepSeek-Reasoner (Token)

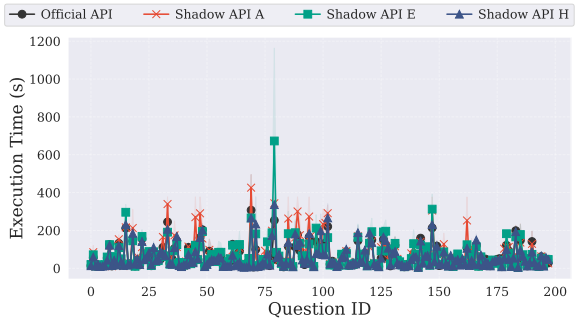
Figure 9: Comparison of inference latency time and token counts on the AIME (Part 2). Solid lines represent mean values, and shaded regions denote the range between the minimum and maximum values across three trials.



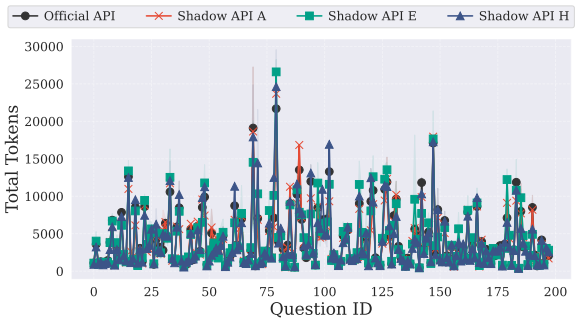
(a) GPQA: GPT-4o-Mini (Time)



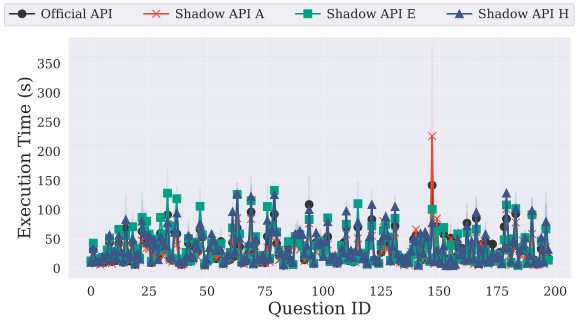
(b) GPQA: GPT-4o-Mini (Token)



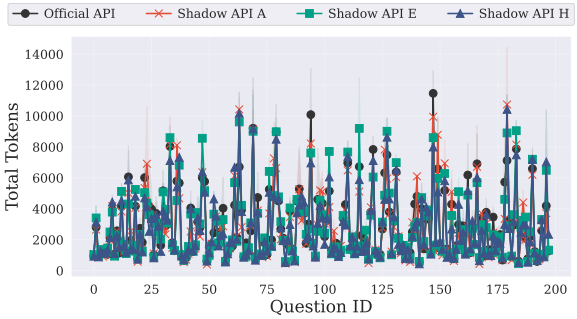
(c) GPQA: GPT-5 (Time)



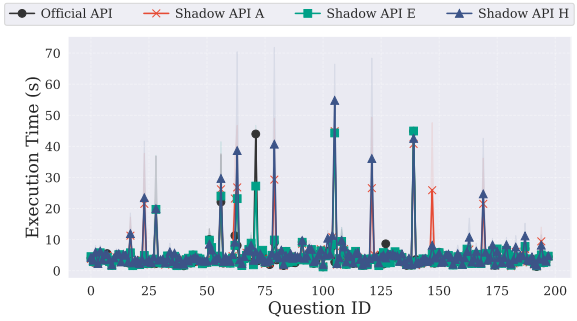
(d) GPQA: GPT-5 (Token)



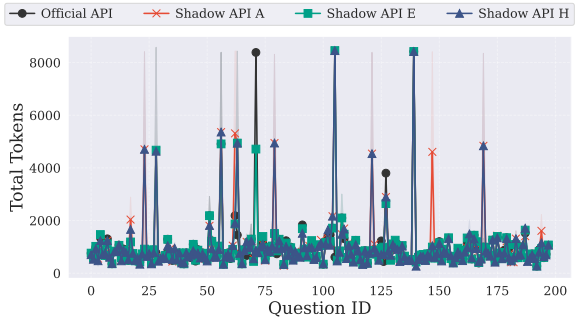
(e) GPQA: GPT-5-Mini (Time)



(f) GPQA: GPT-5-Mini (Token)

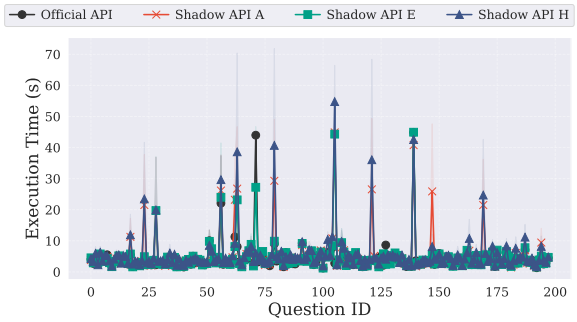


(g) GPQA: Gemini-2.0-Flash (Time)

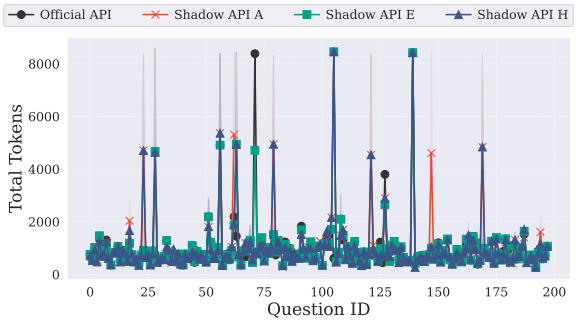


(h) GPQA: Gemini-2.0-Flash (Token)

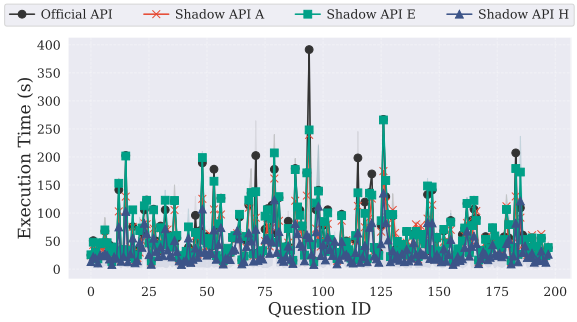
Figure 10: Comparison of inference latency time and token counts on the GPQA (Part 1). Solid lines represent mean values, and shaded regions denote the range between the minimum and maximum values across three trials.



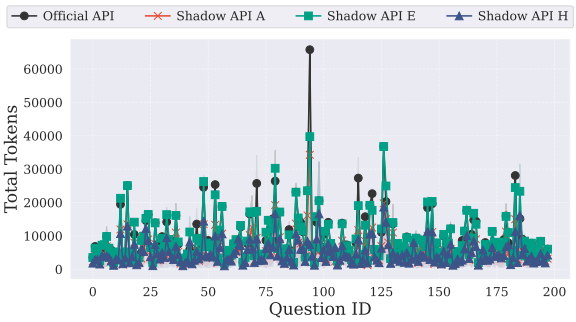
(a) GPQA: Gemini-2.5-Flash (Time)



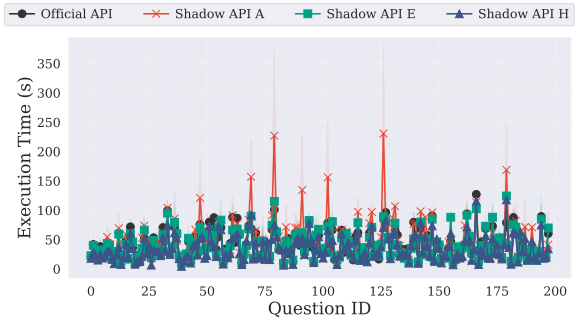
(b) GPQA: Gemini-2.5-Flash (Token)



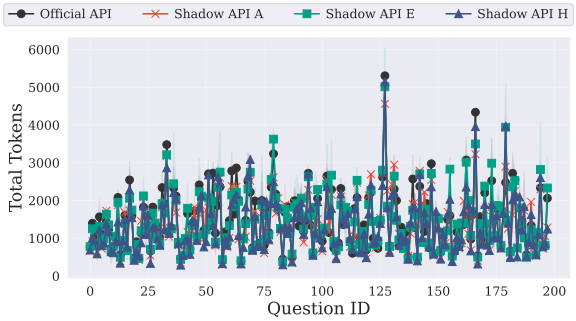
(c) GPQA: Gemini-2.5-Pro (Time)



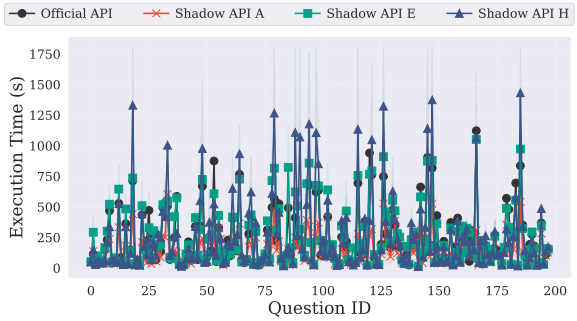
(d) GPQA: Gemini-2.5-Pro (Token)



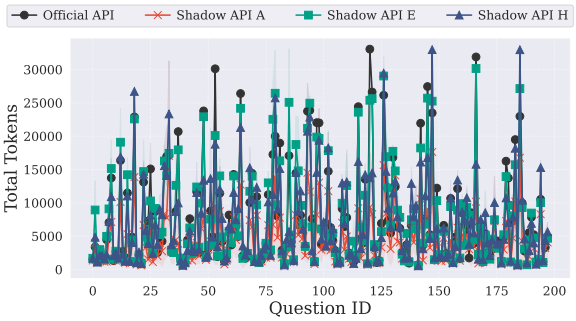
(e) GPQA: DeepSeek-Chat (Time)



(f) GPQA: DeepSeek-Chat (Token)



(g) GPQA: DeepSeek-Reasoner (Time)



(h) GPQA: DeepSeek-Reasoner (Token)

Figure 11: Comparison of inference latency time and token counts on the GPQA (Part 2). Solid lines represent mean values, and shaded regions denote the range between the minimum and maximum values across three trials.

Table 9: List of LLMs used for training and testing LLMmap. The table is divided into the original baseline models and additionally trained models.

#	Version	Provider	Number of Parameters
	Part I: Original Trained Models		
1	ChatGPT-3.5 (gpt-3.5-turbo-0125)	OpenAI	/
2	ChatGPT-4 (gpt-4-turbo-2024-04-09)	OpenAI	/
3	ChatGPT-4o (gpt-4o-2024-05-13)	OpenAI	/
4	Claude 3 Haiku (claude-3-haiku-20240307)	Anthropic	/
5	Claude 3 Opus (claude-3-opus-20240229)	Anthropic	/
6	Claude 3.5 Sonnet (claude-3-5-sonnet-20240620)	Anthropic	/
7	google/gemma-7b-it	Google	7B
8	google/gemma-2b-it	Google	2B
9	google/gemma-1.1-2b-it	Google	2B
10	google/gemma-1.1-7b-it	Google	7B
11	google/gemma-2-9b-it	Google	9B
12	google/gemma-2-27b-it	Google	27B
13	CohereForAI/aya-23-8B	Cohere	8B
14	CohereForAI/aya-23-35B	Cohere	35B
15	Deci/DeciLM-7B-instruct	Deci	7B
16	Qwen/Qwen2-1.5B-Instruct	Qwen	1.5B
17	Qwen/Qwen2-7B-Instruct	Qwen	7B
18	Qwen/Qwen2-72B-Instruct	Qwen	72B
19	gradientai/Llama-3-8B-Instruct-Gradient-1048k	Gradient AI	8B
20	meta-llama/Llama-2-7b-chat-hf	Meta	7B
21	meta-llama/Meta-Llama-3-8B-Instruct	Meta	8B
22	meta-llama/Meta-Llama-3-70B-Instruct	Meta	70B
23	meta-llama/Meta-Llama-3.1-8B-Instruct	Meta	8B
24	meta-llama/Meta-Llama-3.1-70B-Instruct	Meta	70B
25	microsoft/Phi-3-medium-128k-instruct	Microsoft	14B
26	microsoft/Phi-3-medium-4k-instruct	Microsoft	14B
27	microsoft/Phi-3-mini-128k-instruct	Microsoft	3.8B
28	microsoft/Phi-3-mini-4k-instruct	Microsoft	3.8B
29	mistralai/Mistral-7B-Instruct-v0.1	Mistral AI	7B
30	mistralai/Mistral-7B-Instruct-v0.2	Mistral AI	7B
31	mistralai/Mistral-7B-Instruct-v0.3	Mistral AI	7B
32	mistralai/Mixtral-8x7B-Instruct-v0.1	Mistral AI	8x7B
33	nvidia/Llama3-ChatQA-1.5-8B	NVIDIA	8B
34	openchat/openchat-3.6-8b-20240522	OpenChat	8B
35	openchat/openchat_3.5	OpenChat	7B
36	togethercomputer/Llama-2-7B-32K-Instruct	Together AI	7B
37	upstage/SOLAR-10.7B-Instruct-v1.0	Upstage AI	10.7B
38	NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO	Nous Research	8x7B
39	abacusai/Smaug-Llama-3-70B-Instruct	Abacus AI	70B
40	microsoft/Phi-3.5-MoE-instruct	Microsoft	16x3.8B
41	internlm/internlm2_5-7b-chat	InternLM	7B
42	HuggingFaceH4/zephyr-7b-beta	HuggingFace	7B
	Part II: Additional Trained Models		
43	ChatGPT-4o (gpt-4o-mini-2024-07-18)	OpenAI	/
44	ChatGPT-5 (gpt-5-2025-08-07)	OpenAI	/
45	ChatGPT-5-Mini (gpt-5-mini-2025-08-07)	OpenAI	/
46	google/gemini-2.0-flash	Google	/
47	google/gemini-2.5-flash	Google	/
48	google/gemini-2.5-pro	Google	/
49	MiniMaxAI/MiniMax-M2	MiniMaxAI	230B
50	DeepSeek/deepseek-chat	DeepSeek	/
51	DeepSeek/deepseek-reasoner	DeepSeek	/
52	DeepSeek/deepseek-V3.2-exp-chat	DeepSeek	/
53	DeepSeek/DeepSeek-V3-0324	DeepSeek	/
54	Qwen/Qwen2.5-7B-Instruct	Qwen	7B
55	Qwen/Qwen3-VL-32B-Instruct	Qwen	32B
56	Qwen/Qwen3-8B	Qwen	8B
57	DeepSeek/DeepSeek-R1-0528-Qwen3-8B	DeepSeek	8B
58	ZhipuAI/glm-4-9b-chat	ZhipuAI	9B
59	ZhipuAI/glm-Z1-9B-0414	ZhipuAI	9B
60	MoonshotAI/Kimi-K2-Instruct-0905	MoonshotAI	1000B