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

Yage Zhang[†]
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
yage.zhang@cispa.de

Yukun Jiang[†]
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
yukun.jiang@cispa.de

Zeyuan Chen
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
zeyuan.chen@cispa.de

Michael Backes
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
mbackes1978@gmail.com

Xinyue Shen
University of Waterloo
Waterloo, Canada
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
xinyue.shen@uwaterloo.ca

Yang Zhang[*]
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
yangzhang.almo@outlook.com

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 more than 5,900 citations and 58,000 GitHub stars by December 6, 2025. Through multidimensional auditing of three representative shadow APIs across utility, safety, and model verification, we uncover widespread behavioral inconsistency and fingerprint-based evidence consistent with deceptive model claims in a subset of audited endpoints. 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 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. By the time of writing, 4 of the 17 providers have already ceased operations, underscoring the operational volatility of this market. Meanwhile, unverifiable compliance claims and independent model-substitution testing platforms have emerged in the ecosystem, reflecting growing community awareness of this risk.[1]

CCS Concepts

• **Security and privacy → Human and societal aspects of security and privacy**.

Keywords

Large Language Models, Shadow APIs, Third-Party Services, Fingerprint Tests

ACM Reference Format:
Yage Zhang, Yukun Jiang, Zeyuan Chen, Michael Backes, Xinyue Shen, and Yang Zhang. 2026. Real Money, Fake Models: Deceptive Model Claims in Shadow APIs. In *Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS '26), November 15–19, 2026, The Hague, Netherlands.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3830454.3846660

1 Introduction

Frontier large language models (LLMs) have transformed numerous domains, becoming essential infrastructure for scientific research and daily applications [22, 29, 65]. By integrating advanced reasoning and domain knowledge, these models demonstrate enhanced effectiveness in complex tasks [45, 53, 59]. Given 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 [41, 61]. However, official APIs often impose high pricing, payment barriers, and strict geographical restrictions [26]. For instance, the official OpenAI API cannot be accessed directly from several regions, such as China, Russia, and Iran [8, 13]. 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,

[†]Equal contribution.
[*]Corresponding author.

CCS '26, The Hague, Netherlands

ACM ISBN 979-8-4007-2871-6/2026/11
https://doi.org/10.1145/3830454.3846660

[1]The evaluation code is available at https://github.com/TrustAIRLab/ShadowAPI.

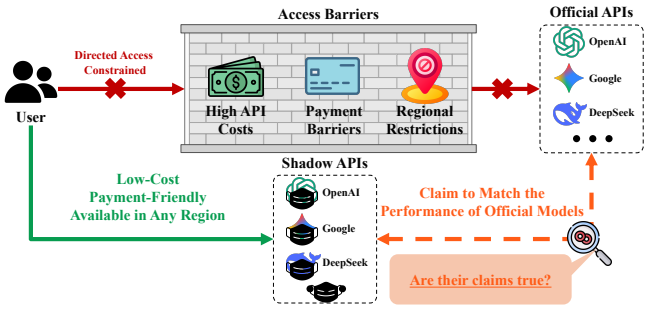


Figure 1: Overview of the shadow API ecosystem.

2025, the shadow APIs we identify have been utilized in 187 academic papers, and the most widely used one alone accounts for more than 5,900 citations and 58,000 GitHub stars. Most of these papers have been accepted by top venues, such as ACL, CVPR, and ICLR (see details in Section 3).

Shadow API use arises within a broader access ecosystem: official frontier-model APIs can be costly or geographically unavailable, while researchers may face resource constraints and publication pressure [14, 60]. We therefore do not treat the use of these services alone as individual misconduct. The security concern arises when results are attributed to a named model without verifiable endpoint provenance, because neither authors nor readers can determine whether the claimed model actually produced the outputs. This motivates our audit of whether shadow endpoints serve the models they claim and our recommendation that endpoint provenance be disclosed and, where feasible, verified.

Recent studies have highlighted the unreliability of API access for open-source models [23, 30, 55], 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. From a security standpoint, the relevant concern is *model provenance integrity*: users cannot directly verify the binding between the advertised model identity and the backend serving their requests. Undocumented backend substitution would violate this binding, whereas operational instability can produce behavioral divergence without establishing an identity mismatch. Both forms of uncertainty can undermine reproducibility and user trust.

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. Shadow APIs also exhibit unpredictable safety divergence from official baselines, with measured harmfulness scores either falling below the official baseline by about 0.23 points or rising to nearly twice the official value. Both directions undermine reproducibility: a lower shadow-endpoint score may understate the risk posed by the official model, whereas a higher score may inflate attack success rates that do not replicate on the official endpoint. Finally, for **RQ3**, we utilize model fingerprinting and output metadata analysis to audit these shadow APIs. We find fingerprint-based evidence consistent with model substitution in a subset of audited endpoints, alongside broader behavioral and metadata inconsistencies. Specifically, across 24 evaluated endpoints, 45.83% fail fingerprint verification, and an additional 12.50% exhibit substantial Euclidean 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, distinguishing inconsistency, substitution, and 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. We did not contact the shadow API providers during the audit or follow-up period because doing so could have altered endpoint behavior or availability.

2 Preliminary

2.1 Background

LLMs have rapidly become the backbone of modern natural language processing (NLP), machine learning (ML), as well as even computer vision (CV) research, and the base for building personal agents such as OpenClaw [38]. 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 [60]. 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 [16]. For instance, Anthropic explicitly disallows sales in unsupported regions [15]. OpenAI likewise warns that accessing or reselling its API from unsupported countries may lead to account suspension [17]. 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

Table 1: Operational distinction between documented distribution channels and shadow APIs.

Criterion	Documented Channel	Shadow API
Upstream route	✓ Publicly documented	✗ Not documented
Credential source	✓ User's own upstream key	✗ Intermediate provider's key
Legal entity disclosed	✓ Verifiable corporation	✗ Typically individual
Registration identifier	✓ Business / ICP registry	✗ Rarely disclosed
Legal documentation	✓ ToS, SLA, privacy policy	✗ Often absent
Regional compliance	✓ Authorized regions only	✗ Bypasses restrictions
Payment accountability	✓ Corporate invoicing	✗ Individual Alipay / WeChat / USDT

accepted papers from regions with restricted access to frontier APIs (e.g., China) [4, 9].

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 [14]. This drives the rise of a commercial shadow market offering discounted access to frontier LLMs without geographical restrictions [6, 12].

Moreover, official terms of service prohibit any form of API key resale or redistribution [47]. In addition, Chinese government regulations require AI services to operate in compliance with applicable laws and administrative requirements.[2] 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

(1) **indirect access** to frontier proprietary models through intermediate providers.
(2) **access provision in restricted regions**, where the service bypasses the official provider's geographic access controls.

To apply these criteria consistently, we classify a service as a shadow API only when it also sells access using credentials or credits issued by the intermediate provider and does not publicly document its upstream route (Table 1). We therefore exclude provider-operated endpoints, publicly documented distribution channels, routing platforms that disclose their upstream routing, and self-hosted gateways that use the user's own credentials. No service was included or excluded based on its measured outputs, and the three providers audited in depth were chosen for popularity, accessibility, and model coverage.

2.3 Security Framing and Scope

Security Property. The relevant security property is *model provenance integrity*: an endpoint sold as access to a named model should preserve the binding between that model identity and the backend serving the request. The risk arises when users cannot verify this binding, causing downstream evaluations to be attributed to a model that may not have produced the outputs.

Interpreting the Evidence. We distinguish measured behavioral divergence from evidence about backend identity. Differences in utility, safety, latency, metadata, or MET outcomes show that an endpoint does not reproduce the official baseline on the measured dimension, but they do not identify the cause or the backend. We report fingerprint-based evidence consistent with model substitution only when the fingerprint matches a specific alternative model. Because black-box fingerprinting is probabilistic, such a match supports but does not prove backend identity. We use *deceptive model claim* only for the subset of endpoints that are commercially offered under one model identity while producing fingerprint evidence consistent with another. This term characterizes the mismatch between the commercial claim and the observed evidence and does not establish operator intent. This claim-centered usage is analogous to regulatory concerns about unsupported AI capability claims [5].

Implications for Security Research. Safety and alignment measurements obtained through an unverified endpoint may characterize a backend other than the model named in the study. A lower measured harmfulness score may understate the risk posed by the named official model, whereas a higher score may overstate attack effectiveness and yield conclusions that do not replicate on the official endpoint. More generally, results attributed to a named model are difficult to reproduce and reliably advance subsequent security/safety research when endpoint provenance is not documented.

Scope of Claims. Ecosystem-level findings concern all 17 identified services and rest on collection, compliance, and metadata evidence (Section 3). Behavioral and identity findings concern only shadow APIs A, E, and H and the 24 endpoints audited directly (Section 4, Section 5). We do not generalize these endpoint-level measurements to the remaining 14 services.

2.4 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 [20, 43, 62]. Pasquini et al. [48] introduce 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. [23] audit model substitution in open-source LLM APIs and evaluate 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 [40] shows that surrogate models can be trained to mimic the outputs of BERT-based NLP APIs, making it difficult for users to distinguish between original APIs and extracted surrogates based on surface-level interactions alone. Recent work examines adjacent risks in the broader LLM deployment ecosystem. Hou et al. [32] conduct an Internet-wide measurement of public-facing LLM deployments, characterizing their prevalence, configurations, and exposure-related security risks. Puthal et al. [49] examine the security and governance risks arising from unsanctioned "shadow AI" use within organizations. These studies focus respectively on insecure public deployments and organizationally ungoverned AI

[2] http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

use. Our audit instead asks whether commercially sold indirect API endpoints actually serve the models they claim. 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) [3, 7]. We use these two venues as seeds because both concentrate LLM-related papers and are recent yet old enough for most associated code repositories to have been publicly released. The collection is not restricted to these two venues, as the endpoint-based expansion below covers repositories from any source. 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 (22.82%) 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.

We include an endpoint when (i) a repository replaces an official provider's base URL with that endpoint while retaining an official model identifier, and (ii) the service satisfies the structural criteria in Section 2.2 and Table 1. We exclude publicly documented cloud distribution channels such as Azure OpenAI, routing platforms that disclose their upstream routing such as OpenRouter, and self-hosted relays that forward a user's own provider credential.

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[3] or journals, such as ACL, CVPR, and ICLR. The most widely used shadow API alone has accumulated more than 5,900 citations, and the associated GitHub repositories have received more than 58,000 stars in total, as shown in Figure 2c. 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.

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 [2] and its derivative, NewAPI [1]. OneAPI is an open-source tool designed for self-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: two services had already ceased operations during our measurements, and four in total had ceased by the time of writing (Section 6.3). In addition, because upstream routing is not transparently documented, our measurements within a bounded observation window cannot establish whether the audited endpoints change backend models over time. Passing our verification checks during one measurement window therefore does not establish continued consistency in subsequent requests or periods. This limited observability constrains users' ability to verify model provenance and assess operational risk.

> **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.

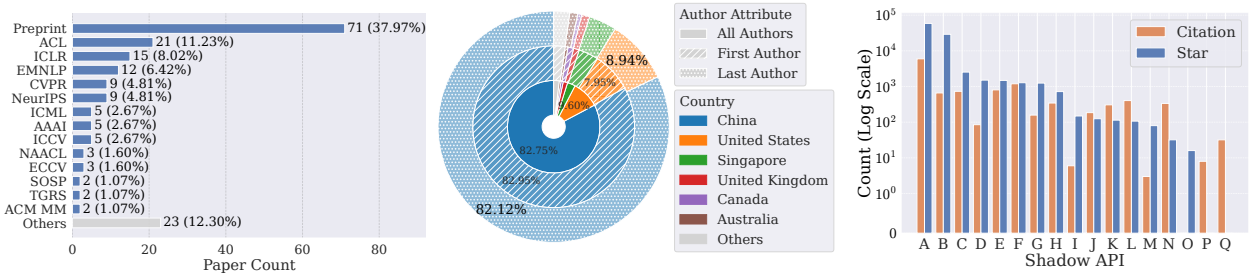
4.1 Experimental Setups

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. 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 [46].
- **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

[3]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.

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 [21]. We assign identifiers to providers (shadow APIs A, B, etc.) based on a descending sort of their associated GitHub star counts, which are summarized in Figure 2c. Specifically, we select shadow APIs A, E, and H based on three criteria: popular with regard to citations and GitHub stars, publicly accessible, and comprehensively covered for models in the three identified model families. 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. Complete model configurations, API parameters, and benchmark settings are provided in our technical report [64].

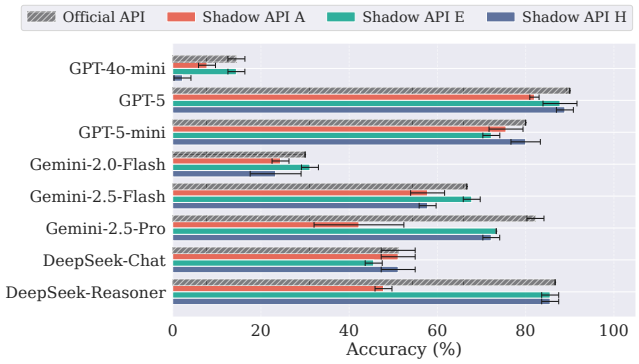
4.2 Utility Evaluation

Benchmarks and Methodology. We assess model utility across scientific and sensitive domains for the evaluated model families, including Gemini 2.5, DeepSeek-V3.2, and Qwen3 [28, 56, 57]. For the science domain, we employ the AIME 2025 benchmark [10], which tests competition-level mathematics, and the GPQA (Diamond) benchmark [51], targeting PhD-level scientific questions. For the sensitive domain, we focus on high-risk fields, i.e., medical and legal fields, where "high-risk" refers to the severity of real-world consequences that incorrect outputs carry for patients and legal parties, rather than to any regulatory classification. In the medical field, following prior work [52], we use the MedQA (USMLE) dataset [39], covering diagnosis, treatment, and medical concepts. In the legal field, we select LegalBench [31], specifically the SCALR task for evaluating legal reasoning and reading comprehension, consistent with [27, 33].

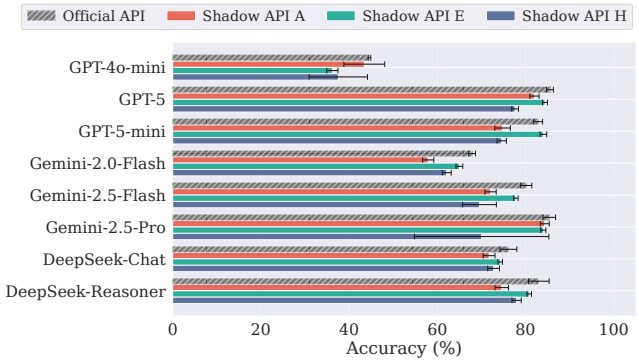
Implementation Details. To ensure fair comparison, we adopt EvalScope prompt templates [11] for the AIME 2025 and GPQA benchmarks, and for the MedQA (USMLE) benchmark, we use the Hulu-Med multiple-choice template [34], and the original task-specific prompts for LegalBench (Scalr) [31].

Science Domain Performance. The official API serves as the reference baseline rather than a strict performance upper bound, as illustrated in Figure 3. Throughout, the quantity of interest is divergence from the official endpoint under identical prompts and decoding configurations, not absolute benchmark accuracy: an endpoint faithfully serving the claimed model should track the official baseline regardless of how high that baseline is. Among the shadow APIs, shadow API E exhibits the strongest consistency, maintaining a minimal average accuracy drop of 2.64% from the official API. In some instances, such as GPT-5-mini on GPQA, it even marginally outperforms the official API by 1.18%, likely reflecting small evaluation or decoding variance. 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.

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 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 2 presents two cases in which all three shadow endpoints return incorrect answers in either all three runs or two of

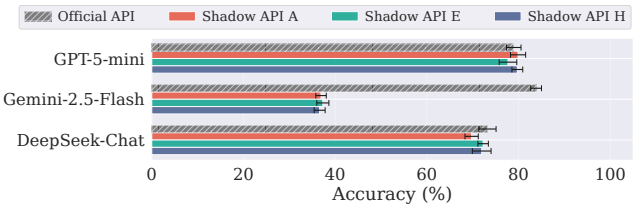


(a) AIME 2025

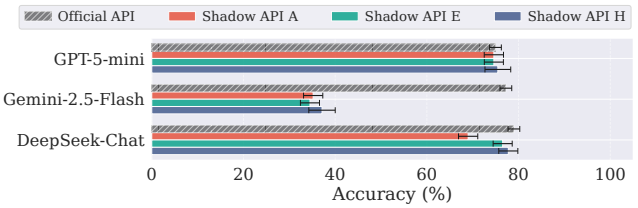


(b) GPQA (Diamond)

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



(a) MedQA (USMLE)



(b) LegalBench (Scalr)

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

Table 2: Medical and legal fields failure response examples.

Benchmark	Question (abridged)	Official API (R1/R2/R3)	Shadow APIs (R1/R2/R3)
MedQA (USMLE) Gemini-2.5-Flash CSV ID 764; Gold: D	A patient in active labor with major bleeding requests transfer from a hospital not equipped for high-risk pregnancy. What is the appropriate next step?	D ✓ / D ✓ / D ✓ Treat and transfer after a written request.	**A, E, H:** E ✗ / E ✗ / E ✗ Transfer only if the benefits outweigh the risks.
LegalBench (Scalr) DeepSeek-Chat CSV ID 147; Gold: 0	Whether nondisclosure of drug adverse-event reports can support a Rule 10b-5 claim when the reports are not statistically significant.	0 ✓ / 3 ✗ / 3 ✗	**A:** 0 ✓ / 1 ✗ / 1 ✗ **E:** 0 ✓ / 3 ✗ / 3 ✗ **H:** 0 ✓ / 4 ✗ / 4 ✗

Entries are ordered as Runs 1/2/3. A checkmark denotes agreement with the benchmark gold answer; a cross denotes disagreement.

the three runs. For Gemini-2.5-flash on MedQA, the three shadow APIs produce answers consistent with the official API in only 49.57% to 49.65% of cases, and the corresponding three-way consistency range on LegalBench is 52.36% to 54.64%. The tightness of this range across three independently operated providers is itself informative. The shadow APIs exhibit nearly identical consistency rates on the same query set, a recurring pattern whose cause cannot be established from performance evidence alone (Section 2.3). GPT-5-mini and DeepSeek-Chat exhibit smaller but still non-trivial consistency gaps, ranging from 80.56% to 96.70%. See our technical report [64] for detailed per-query consistency results. Reliance on shadow APIs for professional guidance therefore poses severe safety risks even when headline accuracy gaps appear moderate.

4.3 Safety Evaluation

Datasets. We consider two widely used benchmarks, JailbreakBench [24] and AdvBench [66], which respectively contain 100 and 520 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 50-item subset provided by [25]. We choose these benchmarks to quantify one dimension of endpoint safety (i.e., response to harmful requests) through black-box API queries without access to model internals. JailbreakBench [24] offers balanced, policy-aligned coverage across ten misuse categories, whereas the representative, redundancy-reduced AdvBench-50 [25, 66] provides a second established inventory and spans 32 harmful categories; using both reduces reliance on any single benchmark's curation. Their wide use in evaluating various jailbreak methods and LLMs [25, 42, 44, 50] supports their relevance as established safety probes.

Implementation Details. We employ four distinct jailbreak attacks to evaluate the safety: GCG [66], Base64 [58], Combination [58], and FlipAttack [42]. For GCG, we use LLaMA3-8B to generate a sample-level suffix and then transfer it to other LLMs. For Base64 and Combination, we follow the settings for Base64 and `combination_1` in [58]. For FlipAttack, we use its well-performed "flip char in sentence" mode.

Metrics. Following prior work [25, 35, 36, 44] such as PAIR, TAP, and JAIL-CON, we introduce a lightweight judge model to output the ***harmfulness score*** and verify whether the generated answer is valid and harmful. Specifically, based on [54], 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.

Safety Evaluation on JailbreakBench. 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 APIs A and E yield substantially lower harmfulness scores than the official endpoint. For Gemini-2.5-flash (Figure 5b), all shadow APIs yield lower measured harmfulness scores 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 instance, shadow APIs A and H generate more (less) harmful content under the attack Combination (FlipAttack).

Safety Evaluation on AdvBench. We observe results on AdvBench similar to those on JailbreakBench (Figure 6), where shadow APIs deviate unpredictably from official endpoints. Specifically, for GPT-5-mini under the Base64 attack (Figure 6a), shadow API A yields a harmfulness score of 0.05, which is 5 × the official API's score of 0.01. Similarly, shadow APIs A, E, and H yield substantially lower harmfulness scores in the Combination attack. For instance, for Gemini-2.5-flash (Figure 6b), all shadow APIs yield lower measured harmfulness scores 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 6c), 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 scores for shadow APIs A and E are comparable to the official endpoints in Base64 and Combination, but lower than the official scores in FlipAttack. 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 through model fingerprinting, distributional testing, controlled validation, and metadata analysis.

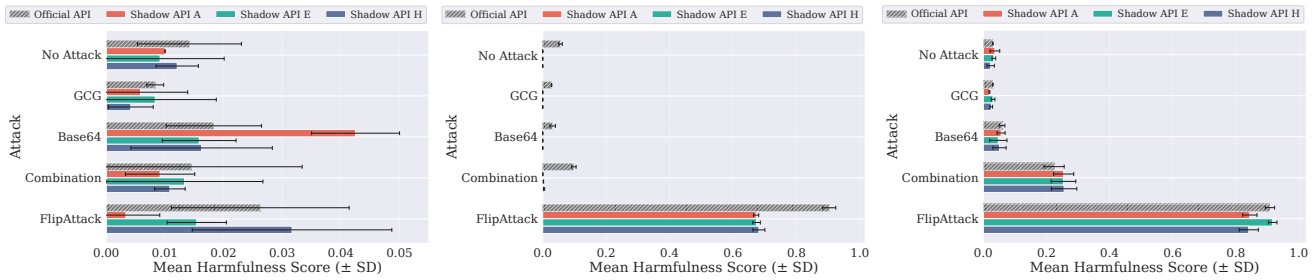
5.1 Fingerprinting-Based Detection

Methodology. To identify the specific LLMs operating behind shadow API service providers, we employ LLMmap [48], an active fingerprinting framework. This framework classifies models by analyzing responses to a curated set of queries and computing the Euclidean distance between model outputs and a reference database. We conduct our fingerprinting experiments following the configuration in [48], 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 reasoning effort.

Results. We observe significant variance in the identity reliability of shadow APIs across the 24 evaluated endpoints, as summarized in Table 3. Specifically, 45.83% of the endpoints fail fingerprint verification, and an additional 12.50% exhibit substantial Euclidean 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 Euclidean 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 Euclidean distances ($D \approx 17.37$–18.04) across all providers.

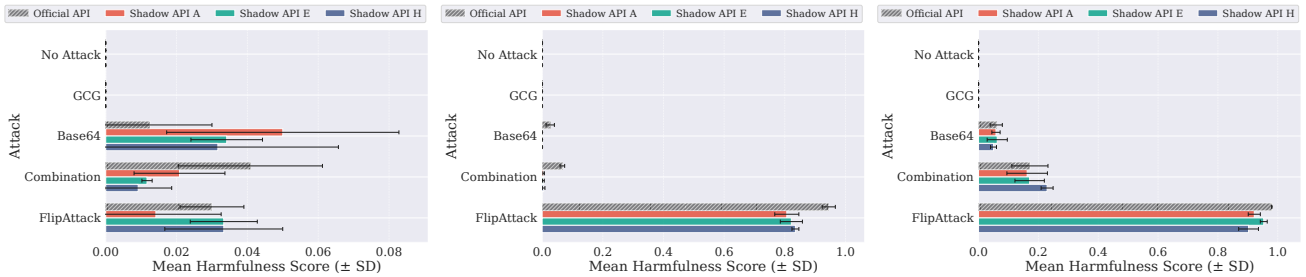
Across the failing endpoints, we observe two recurring model substitution patterns, as supported by the following evidence. First, premium proprietary models exhibit response patterns that more closely align with cheaper open-source alternatives. For instance, the behavior of GPT-4o-mini in shadow API 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.

Capability-Matched Substitution. To examine whether a substituted backend can resemble the claimed model on standard benchmarks, we additionally benchmark the fingerprinted backend in isolation using the same prompt templates and decoding configuration as Section 4. For the GPT-4o-mini→Qwen2.5-7B substitution observed in shadow API H, Qwen2.5-7B achieves 3.33% on AIME 2025 and 42.93% on GPQA, compared to the official GPT-4o-mini's 14.44% and 44.61% on the same benchmarks. The near-parity on GPQA (within 1.68 %) contrasts sharply with an 11.11% gap on AIME 2025, indicating that the matched alternative exhibits a capability profile similar to the claimed model on standard reference tasks, while diverging only under more demanding reasoning settings. This intuitively accounts for two otherwise puzzling observations: the concentration of accuracy drops on reasoning-heavy benchmarks (Figure 3), and the regression finding in Table 9 that identity mismatch alone is not a reliable predictor of accuracy drop.



(a) GPT-5-mini (b) Gemini-2.5-flash (c) DeepSeek-Chat

Figure 5: Safety performance comparison on the JailbreakBench dataset.



(a) GPT-5-mini (b) Gemini-2.5-flash (c) DeepSeek-Chat

Figure 6: Safety performance comparison on the AdvBench dataset.

Table 3: Fingerprinting results via LLMmap matched model and mean Euclidean distance D with standard deviation. Colors denote ratio vs official, Green ($D \leq 1.25\times$ baseline), Yellow ($D > 1.25\times$ 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 $[D: 22.68_{\pm 2.69}]$

5.2 Complementary Verification

To address potential blind spots inherent in relying on a single fingerprinting method, we further employ Model Equality Testing (MET) [30], a two-sample hypothesis test that checks whether shadow API outputs are drawn from the same distribution as those of the official model. When the null hypothesis of distributional equality is rejected at $\alpha = 0.05$ (*Reject* = True), the shadow API's outputs are statistically distinguishable from the official model's, providing independent evidence of identity inconsistency. We evaluate three shadow API providers (A, E, H) across four benchmarks (AIME 2025, GPQA, AdvBench, and JailbreakBench), covering eight model configurations in total. The full results are reported in Table 4. Overall, MET and LLMmap provide complementary rather than

Table 4: MET results across shadow API providers. *Reject* = True indicates the null hypothesis of distributional equality is rejected at $\alpha = 0.05$.

Benchmark	Model	Shadow API A	Shadow API E	Shadow API H
AIME 2025	GPT-4o-mini	**Reject**	Pass	**Reject**
	GPT-5	**Reject**	**Reject**	Pass
	GPT-5-mini	Pass	Pass	**Reject**
	Gemini-2.0-flash	**Reject**	Pass	Pass
	Gemini-2.5-flash	Pass	Pass	Pass
	Gemini-2.5-pro	Pass	Pass	Pass
	DeepSeek-Chat	**Reject**	**Reject**	**Reject**
	DeepSeek-Reasoner	**Reject**	Pass	**Reject**
GPQA	GPT-4o-mini	**Reject**	Pass	**Reject**
	GPT-5	**Reject**	**Reject**	**Reject**
	GPT-5-mini	**Reject**	Pass	Pass
	Gemini-2.0-flash	**Reject**	**Reject**	Pass
	Gemini-2.5-flash	**Reject**	Pass	**Reject**
	Gemini-2.5-pro	**Reject**	Pass	Pass
	DeepSeek-Chat	**Reject**	**Reject**	**Reject**
	DeepSeek-Reasoner	**Reject**	Pass	**Reject**
AdvBench	GPT-5-mini	**Reject**	Pass	**Reject**
	Gemini-2.5-flash	**Reject**	**Reject**	**Reject**
	DeepSeek-Chat	**Reject**	Pass	**Reject**
JailbreakBench	GPT-5-mini	**Reject**	Pass	Pass
	Gemini-2.5-flash	**Reject**	**Reject**	**Reject**
	DeepSeek-Chat	**Reject**	Pass	**Reject**

identical evidence. Because LLMmap gives endpoint-level identity signals while MET gives benchmark-specific distributional tests, we compare them qualitatively. The MET results broadly corroborate the fingerprinting findings and further reveal behavioral divergence beyond endpoint identity.

For instance, GPT-4o-mini and GPT-5 in shadow API A are both flagged as statistically distinct from official outputs across AIME 2025 and GPQA, consistent with the Euclidean distance deviations observed via LLMmap. DeepSeek-Chat is rejected across nearly all provider benchmark combinations, reinforcing the identity inconsistency identified in Table 3. The Gemini-2.5-pro results again stand out as an exception: MET fails to reject the null hypothesis across all three providers on AIME 2025. This indicates that, under this benchmark and sample size, we do not observe distributional evidence that the shadow outputs differ from the official endpoint. It should not be interpreted as certifying faithful serving, since MET only tests distributional equality on the evaluated samples.

5.3 Validation on Controlled Ground Truth

The audits in Section 5 and Section 4 apply verification methods to production shadow endpoints whose true backends are not disclosed. To calibrate these methods under known ground truth, we construct a local testbed where honest endpoints serve the claimed model and substituting endpoints silently serve an alternative backend reflecting the substitution patterns observed in Table 3. This controlled setting provides an empirical sanity check for the false positive and false negative behavior of our audit pipeline.

We construct six endpoints spanning the three model families studied in this paper, with three honest and three substituting configurations, as summarized in Table 5. Each endpoint serves N=100 queries for LLMmap evaluation, and an independent set of 600 requests is issued for MET evaluation. The substituting configurations cover three substitution modes: cross-family model substitution (GPT-5 served by GLM-4-9B), within-family version mismatch (Gemini-2.0-flash served by Gemini-2.5-flash), and capability downgrading (DeepSeek-Reasoner served by the non-thinking DeepSeek-Chat).

Table 5: Controlled validation testbed.

Type	Claimed Model	Backend	Substitution Mode
Honest	GPT-5	GPT-5	None
Honest	Gemini-2.0-flash	Gemini-2.0-flash	None
Honest	DeepSeek-Reasoner	DeepSeek-Reasoner	None
Substituting	GPT-5	GLM-4-9B	Cross-family
Substituting	Gemini-2.0-flash	Gemini-2.5-flash	Version mismatch
Substituting	DeepSeek-Reasoner	DeepSeek-Chat	Capability downgrade

Table 6: Decision-level performance of LLMmap and MET on the controlled validation testbed.

	LLMmap			MET		
Model Family	**Acc**	**FPR**	**FNR**	**Acc**	**FPR**	**FNR**
GPT family	96.00%	3.00%	5.00%	88.50%	5.00%	18.00%
Gemini family	96.00%	2.00%	6.00%	87.00%	5.00%	21.00%
DeepSeek family	96.00%	4.00%	4.00%	89.50%	4.00%	17.00%
Overall	96.00%	3.00%	5.00%	88.33%	4.67%	18.67%

Table 6 reports decision-level performance on this controlled testbed. For LLMmap, one decision corresponds to running the fingerprinting procedure on a sampled query set and comparing the predicted backend with the known ground-truth backend. For MET, one decision corresponds to testing whether responses from the served backend and the claimed official model are drawn from the same distribution. A false positive denotes an honest endpoint incorrectly flagged as inconsistent, while a false negative denotes a substituting endpoint incorrectly accepted as faithful.

LLMmap achieves 96.00% overall accuracy with a 3.00% false positive rate and a 5.00% false negative rate. MET achieves 88.33% accuracy with a 4.67% false positive rate and an 18.67% false negative rate. Performance is consistent across the three model families, with no single family diverging by more than 1.3% from the overall accuracy (Table 6). Both methods correctly flag the three controlled substitution modes, providing a sanity check that the production findings are not solely artifacts of the verification methods.

The elevated FNR of MET (18.7%) reflects its operating regime. MET detects distributional shift between outputs rather than exact model identity, so it may fail to reject when a substituted backend produces outputs statistically close to the claimed model, for instance under small version mismatches within the same family. Thus, MET and LLMmap capture different failure modes. These validation results provide an empirical sanity check for the audit pipeline and increase confidence that the fingerprint failures and distributional rejections observed on production shadow APIs reflect genuine inconsistencies rather than method artifacts alone.

5.4 Meta Information Analysis

Beyond model-level identity, the routing behavior of shadow APIs leaves observable traces in inference-level metadata. Official APIs typically exhibit consistent latency and token counts when the

same question is issued repeatedly, whereas shadow APIs frequently exhibit irregular spikes consistent with dynamic backend switching or upstream retries. To quantify this, for each of the 24 shadow endpoints on each of the two benchmarks (AIME 2025 and GPQA), we compute the standard deviation (SD) ratio $r = \mathrm{SD}_{\mathrm{shadow}}/\mathrm{SD}_{\mathrm{official}}$ separately for inference latency and total token counts, yielding n=48 endpoint-benchmark pairs per dimension.

Table 7 and Table 8 report per-endpoint SD values with color-coded ratio bands. Inference latency exhibits pronounced instability, with 64.58% of pairs exceeding $r = 1.20$ and 35.42% exceeding $r = 2.00$. Every case with $r > 2.00$ on latency falls within the Gemini or DeepSeek families, while no GPT-series endpoint reaches this threshold, suggesting that upstream routing chains for non-OpenAI models are substantially less stable. Token-count ratios are more evenly distributed but still reveal 20.83% of pairs with $r > 2.00$, again concentrated in Gemini and DeepSeek. At the lower extreme, 20.83% of token-count pairs fall below $r = 0.80$, dominated by GPT-series endpoints on AIME 2025, a pattern consistent with upstream truncation or fixed-length response caps rather than faithful replication of the official endpoint.

RQ3 Take-Aways: Fingerprinting yields specific alternative-model matches for a subset of audited endpoints, providing evidence consistent with model substitution. More broadly, 45.83% of endpoints fail fingerprint verification, MET rejects distributional equality for the majority of provider-benchmark combinations, and over one-third of endpoint-benchmark pairs exhibit latency volatility exceeding 2.00×. These broader signals establish endpoint inconsistency but do not by themselves identify the backend. When an endpoint is commercially offered under one model identity while its fingerprint matches another, the evidence establishes a deceptive model claim.

6 Discussion

Building on the empirical results from the preceding sections, we synthesize findings across evaluation dimensions (Section 6.1), quantify the economic incentives underlying deceptive model claims (Section 6.2), and report observed ecosystem changes during and after our study window(Section 6.3).

6.1 Joint Analysis

Based on the combined results from performance evaluation (Section 4) and model verification (Section 5), we find that model identity and observed performance do not always align.

When Identity Predicts Performance. Most endpoints follow an intuitive pattern in which matching identity coincides with consistent behavior and identity mismatch coincides with degradation. GPT-5-mini in shadow API E illustrates the former: fingerprinting matches the claimed model, and sensitive-domain accuracy remains close to the official endpoint. DeepSeek-Reasoner in shadow API A and GPT-4o-mini in shadow API H illustrate the latter, where fingerprint mismatch is accompanied by sharp reasoning collapse on AIME 2025.

When Identity and Performance Diverge. Two counterintuitive patterns violate this alignment. In one direction, identity mismatch 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 2025 accuracy declines by 2.22 to 7.78 percentage points. This gap is substantially smaller than the capability differential between a frontier proprietary model and a 9B open-source backend would predict. A substituted backend may nevertheless track the claimed model on standard reference tasks (Section 5), so degradation surfaces most clearly under reasoning-heavy or strict-correctness conditions. In the other direction, matching identity does not guarantee faithful behavior. Gemini-2.5-flash fingerprints correctly across shadow APIs A, E, and H with Euclidean distances close to the official APIs, yet accuracy in sensitive domains drops by roughly 47% uniformly across all three providers. A plausible conjecture is that these providers serve the authentic model with its thinking budget disabled or sharply reduced to cut serving cost, which would leave fingerprints intact while selectively collapsing performance on reasoning-heavy queries.

To quantify these relationships, we fit an OLS regression predicting accuracy drop from Euclidean distance, price ratio, identity mismatch, and a reasoning-model flag across all 24 endpoints. The full specification is reported in Table 9.

Price Does Not Signal Quality. As shown in Table 9, price ratio has no predictive power for accuracy drop ($\beta = -1.447$, $p = 0.490$), and this null result persists when the regression is restricted to the subset of endpoints passing fingerprint verification. This directly refutes the intuition that more expensive shadow APIs deliver more faithful service. Users cannot protect themselves from substitution-induced quality loss by selecting higher-priced providers, and price cannot serve as even a weak proxy for backend fidelity.

Fingerprint Fidelity Does Not Imply Behavioral Fidelity. The low overall fit ($R^2 = 0.15$) is driven by the Gemini-2.5-flash anomaly documented above. Excluding these three endpoints, the fit improves to $R^2 = 0.30$ and identity mismatch becomes marginally significant ($\beta = 7.72$, $p = 0.05$), indicating that identity substitution is a meaningful driver of accuracy drop once this anomaly is set aside. The anomaly itself, however, carries the stronger message: identity-consistent serving does not guarantee behavioral fidelity. A provider can route requests to the claimed model and still degrade output quality through misconfigured inference parameters, context-window truncation, or silent prompt preprocessing, mechanisms that Euclidean-distance-based fingerprinting is not designed to detect. Verification pipelines relying solely on identity checks, therefore, cannot certify that a shadow API faithfully reproduces the official endpoint's behavior, and behavioral benchmarking remains a necessary complement to identity-level auditing.

Consolidated Endpoints Evidence. To provide a unified view, Table 10 consolidates the four independent evidence dimensions evaluated in this work for each of the 24 shadow endpoints: fingerprint identity (FP), distributional equality testing (MET), performance inconsistency (Perf), and metadata instability (Meta, latency SD ratio r>2.00 on either AIME or GPQA). Each endpoint is assigned an aggregate risk level: **High** (≥3 flags), **Medium** (1–2 flags), or **Low** (0 flags). Of the 24 endpoints, 11 (45.83%) are classified as High risk, 9 (37.50%) as Medium, and only 4 (16.67%) as Low. Shadow API A contains no Low-risk endpoint, while three of the four Low-risk

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

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×), Green (0.8× to 1.2×), Red (> 1.2×).

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

Table 9: OLS regression predicting accuracy drop (Δacc).

Variable	β	SE	t	p
Euclidean Distance	−0.39	1.41	−0.27	0.79
Price Ratio	−1.45	2.06	−0.70	0.49
Identity Mismatch	0.22	7.25	0.03	0.98
Reasoning Model	10.09	7.54	1.34	0.20
$R^2 = 0.15$, $n = 24$				

endpoints belong to shadow API E, confirming that the distribution of flags differs across the three audited providers.

Joint Analysis Take-Aways: Price ratio has no predictive power for accuracy degradation, so pricing cannot serve as a quality signal for backend fidelity. Identity-consistent serving does not guarantee behavioral fidelity: a shadow API can pass fingerprint verification and still degrade output quality through non-identity mechanisms.

6.2 Economic Incentives and User Losses

We observe three recurring pricing patterns in the relationship between advertised model identities and fingerprinted backends, summarized in Table 11. Each instantiates the transaction structure of a deceptive model claim, a substituted backend sold and priced under the claimed model name. In the information premium scheme,

Table 10: Consolidated per-endpoint evidence summary across four independent dimensions.

		Shadow API A					Shadow API E					Shadow API H				
Family	**Model**	**FP**	**MET**	**Perf**	**Meta**	**Risk**	**FP**	**MET**	**Perf**	**Meta**	**Risk**	**FP**	**MET**	**Perf**	**Meta**	**Risk**
GPT	GPT-4o-mini	✓	✓	–	–	Med	–	–	–	–	Low	✓	✓	✓	–	High
	GPT-5	✓	✓	✓	–	High	✓	✓	✓	–	High	–	–	–	–	Low
	GPT-5-mini	✓	✓	–	–	Med	–	–	–	–	Low	✓	✓	–	–	Med
Gemini	Gemini-2.0-flash	✓	✓	✓	✓	High	✓	–	✓	✓	High	–	–	–	✓	Med
	Gemini-2.5-flash	✓	✓	✓	✓	High	–	✓	✓	–	Med	–	✓	✓	✓	High
	Gemini-2.5-pro	–	–	✓	✓	Med	–	–	–	✓	Med	–	–	–	✓	Med
DeepSeek	DeepSeek-Chat	✓	✓	✓	✓	High	✓	✓	–	–	Med	✓	✓	–	✓	High
	DeepSeek-Reasoner	✓	✓	✓	✓	High	–	–	–	–	Low	✓	✓	✓	✓	High
Per-provider summary		5 High, 3 Med, 0 Low					2 High, 3 Med, 3 Low					4 High, 3 Med, 1 Low				

providers charge a premium rate while the endpoint fingerprints as a cheaper model of similar or newer vintage (e.g., API A advertises Gemini-2.0-flash but fingerprints as Gemini-2.5-flash at a 7.1–7.25× price ratio). In the discount-substitution scheme, providers charge at the official rate while the endpoint fingerprints as a low-cost open-source backend (e.g., API A advertises GPT-5 at parity pricing but fingerprints as GLM-4-9B). In the resale markup scheme, providers apply a modest surcharge while the endpoint fingerprints as a different backend, as illustrated by API H, which charges 1.09× the official rate for GPT-4o-mini while its responses fingerprint as Qwen2.5-7B.

Value Delivered vs. Cost Incurred.

As a descriptive supplement, we report an aggregate cost snapshot recorded during an earlier exploratory measurement of GPT-5 on GPQA, using official GPT-5 pricing ($1.25/$10.00 per 1M prompt/completion tokens). Table 12 reports the aggregate estimates retained from this measurement as descriptive context. We do not rely on them for our primary endpoint-fidelity conclusions.

Aggregate Research Costs. Beyond per-query losses, shadow API deception imposes broader costs on the research community. Our dataset covers 187 papers whose LLM-based pipelines rely on shadow API providers. As an illustrative scenario rather than an empirical estimate, if 30% of these papers were to require re-execution upon detecting identity inconsistencies ($n \approx 56$), the aggregate direct cost, comprising API re-runs ($50–$500 per paper) and researcher time (approximately 40 h × $50/h = $2,000 per paper), would range from $115,000 to $140,000. This estimate excludes downstream reproducibility costs associated with the more than 5,900 cumulative citations to these papers, where silent model substitution may propagate into dependent experimental results without any visible error signal.

Because providers observe their routing choices while users observe only a model name and price, this market exhibits the information asymmetry associated with a market-for-lemons dynamic [18, 19].

6.3 Observed Ecosystem Changes

During and after our study window, we observe two categories of ecosystem changes. We did not notify shadow API providers, and we do not attribute either category to our disclosure.

Provider Cessation. By the time of writing, four of the 17 identified shadow API providers ceased operations entirely. They strand users who had ongoing research pipelines dependent on these endpoints, further illustrating the operational volatility documented in Section 3.

Compliance Adoption and Testing Platforms. Beyond the providers in our collection, third-party LLM API services in the broader market are changing by introducing compliance-related claims on their platforms. For instance, several providers published "model verification passed" badges on their pricing pages. Concurrently, some testing platforms dedicated to detecting model substitution in third-party LLM API services have emerged (e.g., https://hvoy.ai/ and https://cctest.ai/). These platforms offer automated checks that attempt to verify whether an LLM API endpoint serves the claimed model. Their emergence indicates a growing awareness of the risks associated with LLMs provided by third-party LLM API providers.

7 Suggestions

Our findings reveal that model substitution is widespread among the audited endpoints and technically detectable. To safeguard the integrity of LLM-based research, we propose a concrete two-workflow framework for auditors and researchers respectively. The fundamental mitigation remains direct official access, and we recommend it wherever available. Since prohibition alone is not actionable under regional restrictions and prohibitive pricing (Section 2.3), a shadow endpoint should instead be treated as an untrusted instrument: verified before use, monitored during use, and disclosed at publication.

Auditor Verification Protocol. We recommend a four-stage verification pipeline in which any flagged stage should escalate immediately to avoidance of the endpoint. First, the endpoint should be queried with at least 24 LLMmap probes, flag it if the Euclidean distance exceeds 1.25× the official baseline, or if the top-1 identified model does not match the claimed identity. Second, MET [30] should be applied using a permutation-based p-value with a prespecified significance level (e.g., $\alpha = 0.05$), and the endpoint should be flagged if the null hypothesis of distributional equality is rejected. Third, at least three independent sessions should be run on a held-out benchmark, flag the endpoint if the accuracy standard deviation exceeds 5 % or if the latency coefficient of variation exceeds 0.15, as excessive variance is indicative of unstable backend routing or dynamic model substitution. Fourth, ICP registration and legal entity disclosure should be verified against the criteria. An

Table 11: Three observed pricing patterns associated with deceptive model claims.

Mechanism	Example	Price Ratio	Deception Evidence
Information Premium	API A — Gemini-2.0-flash	7.10 × / 7.25 ×	Fingerprinted as Gemini-2.5-flash
Discount-Substitution	API A — GPT-5	1.00 ×	Fingerprinted as GLM-4-9B (cheap-for-premium swap)
Resale Markup	API H — GPT-4o-mini	1.09 ×	Fingerprinted as Qwen2.5-7B

Table 12: Historical aggregate cost snapshot for GPT-5 on GPQA under official-token pricing.

API	Equiv. Value	Relative to Off.	Value Shortfall
Official	\$14.84	1.00×	–
Shadow API A	\$5.70	0.38×	\$9.14
Shadow API E	\$5.35	0.36×	\$9.49
Shadow API H	\$7.77	0.52×	\$7.07

endpoint operated by an unregistered individual without verifiable provenance carries elevated legal and operational risk.

Researcher Pre-Registration Checklist. For any study relying on LLM API queries, researchers should complete and publicly report the following prior to data collection. The full endpoint URL, claimed model version, date of access, and pricing tier should be recorded and included as a footnote or within the experimental setup section, consistent with emerging reproducibility norms. Before any experimental queries are issued, the endpoint should be confirmed to pass all four stages of the auditor protocol above. Results from endpoints that fail any stage should not be reported as representative of the claimed model. Because providers switch upstream backends dynamically (Section 3), a one-time check is not sufficient: lightweight probes (e.g., a fixed subset of LLMmap queries) should be interleaved daily or every few thousand requests. Raw responses and per-request metadata (endpoint URL, model identifier, latency, token counts, and logprobs) should be archived and released as artifacts, so provenance can be re-examined post hoc. Finally, per-run accuracy over at least three independent runs, the LLMmap fingerprint Euclidean distance, and the MET p-value should be reported alongside experimental results, enabling readers and reviewers to assess backend reliability and contextualize reported performance accordingly.

Community-Level Actions. Beyond individual practice, we call on the broader research community to adopt structural safeguards. Conference organizers and program chairs should update reviewer guidelines to flag undisclosed or unverified third-party API endpoints as a reproducibility risk, treating such usage analogously to unverified dataset provenance. Official model providers can further reduce shadow market demand through channels compatible with export-control constraints: academic pricing and sponsored research access, cross-region collaborations in which co-authors in supported regions run the official queries, and lightweight verification endpoints for independent identity checks. Where frontier capability is not strictly required, openly released models with self-hosted weights or TEE-attested serving [23] eliminate the provenance question, and the emerging substitution-testing platforms (Section 6.3) could grow into community infrastructure that continuously publishes endpoint-fidelity reports.

8 Limitations

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

Temporal Scope and Market Volatility. Our study reflects a deliberately bounded snapshot within a specific time window (September to December 2025). Shadow API providers frequently switch upstream model sources, alter routing strategies, or cease operations without notice, so behavioral characteristics may drift beyond our observation window. However, the core risk signals we document, non-transparent model substitution, capability downgrading, and safety behavior divergence, reflect structural properties of the service category rather than artifacts of any single measurement period. We intentionally adopt a snapshot design rather than continuous monitoring to ensure experimental consistency and to avoid inadvertently amplifying the shadow API ecosystem through sustained engagement.

Detection Coverage and Ground Truth. In the absence of ground truth regarding backend infrastructure, our investigation relies on established auditing frameworks (LLMmap and MET) alongside performance and metadata analysis. While we detect model identity mismatches with high confidence, we cannot definitively distinguish all fine-grained implementation details (e.g., silent prompt preprocessing or context-window truncation). Nor can black-box verification establish the nature of a substituted backend, i.e., whether it is distilled, fine-tuned, quantized, or version-shifted [23, 30]. Building on model-extraction results for BERT-based NLP APIs [40], a sufficiently faithful surrogate could in principle evade both LLMmap and MET, since output-based detection is weakest against substitutions that closely track the reference distribution [23]. Passing both tests therefore indicates absence of evidence of substitution rather than proof of authenticity. Moreover, official APIs of proprietary models may undergo undisclosed fine-tuning. We mitigate this by reporting variance across three independent runs per experiment. We also recognize the potential to analyze shadow API differences from additional perspectives, such as bias and hallucination [37, 63].

Coverage of Providers and Model Families. Our 17 providers, collected via academic citations and GitHub stars, may not represent the full shadow API market. Additional services may employ distinct deceptive strategies. Similarly, we focus on three representative model families. We believe our audit pipeline generalizes to broader model coverage and leave this to future work.

9 Conclusion

In this work, we present the first systematic audit of the shadow API, a widely used but unverified access for frontier LLMs. Through a comprehensive evaluation across providers and benchmarks, together with behavioral fingerprinting, we demonstrate 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 a subset of audited shadow endpoints and demonstrate that the audited shadow endpoints cannot be treated as reliable official models.

References

[1] 2023. New API. https://github.com/QuantumNous/new-api.
[2] 2023. One API. https://github.com/songquanpeng/one-api.
[3] 2024. ACL 2024 Accepted Main Conference Papers. https://2024.aclweb.org/program/main_conference_papers/.
[4] 2024. CVPR'24 in Numbers. https://latticeflow.ai/news/cvpr24-in-numbers.
[5] 2024. FTC Announces Crackdown on Deceptive AI Claims and Schemes. https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes.
[6] 2024. Hackers Are Stealing GenAI Credentials, So What Sensitive Company Data Are They Getting their Hands On? https://www.esentire.com/blog/hackers-are-stealing-genai-credentials-so-what-sensitive-company-data-are-they-getting-their-hands-on.
[7] 2024. ICLR 2024 Papers. https://iclr.cc/virtual/2024/papers.html.
[8] 2024. OpenAI Taking Steps to Block China's Access to Its AI Tools. https://www.bloomberg.com/news/articles/2024-06-25/openai-warns-it-will-block-access-to-ai-tools-from-china.
[9] 2025. AAAI-26 Review Process Update: Scale, Integrity Measures, and Pathways to Sustainability. https://aaai.org/conference/aaai/aaai-26/review-process-update/.
[10] 2025. AIME Problems and Solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.
[11] 2025. EvalScope: LLM Benchmarks. https://evalscope.readthedocs.io/en/v1.0.0/get_started/supported_dataset/llm.html.
[12] 2025. How to Use Banned US Models in China. https://www.chinatalk.media/p/the-grey-market-for-american-llms.
[13] 2025. Supported countries and territories. https://platform.openai.com/docs/supported-countries.
[14] 2025. The Unsustainable Economics of LLM APIs: Understanding the Coming Price Realignment. https://www.linkedin.com/pulse/unsustainable-economics-llm-apis-understanding-coming-vaidheeswaran-h7hpc/.
[15] 2025. Updating restrictions of sales to unsupported regions. https://www.anthropic.com/news/updating-restrictions-of-sales-to-unsupported-regions.
[16] 2026. Available regions for Google AI Studio and Gemini API. https://ai.google.dev/gemini-api/docs/available-regions.
[17] 2026. ChatGPT and API services in unsupported countries and territories. https://help.openai.com/en/articles/9131992-chatgpt-and-api-services-in-unsupported-countries-and-territories.
[18] George A. Akerlof. 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics* (1970).
[19] Ross J. Anderson. 2001. Why Information Security is Hard-An Economic Perspective. In *Annual Computer Security Applications Conference (ACSAC)*. IEEE Computer Society, 358–365.
[20] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *International Conference on Learning Representations (ICLR)*.
[21] Hudson Borges, André C. Hora, and Marco Túlio Valente. 2016. Understanding the Factors That Impact the Popularity of GitHub Repositories. In *International Conference on Software Maintenance and Evolution (ICSME)*. 334–344.
[22] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR abs/2303.12712* (2023).
[23] 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* (2025).
[24] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, 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)*. NeurIPS, 55005–55029.
[25] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2025. Jailbreaking Black Box Large Language Models in Twenty Queries. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE.
[26] 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* (2023).
[27] 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. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*. ACL, 3275–3293.
[28] DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *CoRR abs/2512.02556* (2025).
[29] 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* (2025).
[30] Irena Gao, Percy Liang, and Carlos Guestrin. 2025. Model Equality Testing: Which Model is this API Serving?. In *International Conference on Learning Representations (ICLR)*.
[31] 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, Jessica Wu, Joe Nudell, Joel Niklaus, John J. Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael A. Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.
[32] Xinyi Hou, Jiahao Han, Yanjie Zhao, and Haoyu Wang. 2025. Unveiling the Landscape of LLM Deployment in the Wild: An Empirical Study. *CoRR abs/2505.02502* (2025).
[33] 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. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 13759–13781.
[34] 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, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. 2025. Hulu-Med: A Transparent Generalist Model towards Holistic Medical Vision-Language Understanding. *CoRR abs/2510.08668* (2025).
[35] Yukun Jiang, Hai Huang, Mingjie Li, Yage Zhang, Michael Backes, and Yang Zhang. 2026. Sparse Models, Sparse Safety: Unsafe Routes in Mixture-of-Experts LLMs. *CoRR abs/2602.08621* (2026).
[36] Yukun Jiang, Mingjie Li, Michael Backes, and Yang Zhang. 2025. Adjacent Words, Divergent Intents: Jailbreaking Large Language Models via Task Concurrency. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.
[37] Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, and Yang Zhang. 2024. ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 12814–12845.
[38] Yukun Jiang, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. 2026. "Humans welcome to observe": A First Look at the Agent Social Network Moltbook. *CoRR abs/2602.10127* (2026).
[39] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *CoRR abs/2009.13081* (2020).
[40] Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. 2020. Thieves on Sesame Street! Model Extraction of BERT-based APIs. In *International Conference on Learning Representations (ICLR)*.
[41] 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, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. *CoRR abs/2211.09110* (2022).
[42] Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, and Bryan Hooi. 2024. FlipAttack: Jailbreak LLMs via Flipping. *CoRR abs/2410.02832* (2024).
[43] Hope McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. 2024. Your Large Language Models Are Leaving Fingerprints. *CoRR abs/2405.14057* (2024).

[44] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2023. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. *CoRR abs/2312.02119* (2023).
[45] OpenAI. 2023. GPT-4 Technical Report. *CoRR abs/2303.08774* (2023).
[46] OpenAI. 2025. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf.
[47] OpenAI. 2025. OpenAI Services Agreement. https://openai.com/policies/may-2025-business-terms/.
[48] Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. 2025. {LLMmap}: Fingerprinting for Large Language Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 299–318.
[49] Deepak Puthal, Niranjan Kumar Ray, and Saraju P. Mohanty. 2025. Shadow AI and the Expanding Attack Surface: Risks, Threats, and Countermeasures. In *OITS International Conference on Information Technology (OCIT)*.
[50] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR abs/2403.05530* (2024).
[51] 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* (2023).
[52] Yanjun Shao, Xiangru Tang, Jiwoong Sohn, Jiapeng Chen, Yuxuan Liao, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, Arman Cohan, and Mark Gerstein. 2025. MedicalAgentsBench for Complex Medical Reasoning: Comparing Internalized Reasoning Models versus Externalized Agent-based Frameworks. *CoRR abs/2503.07459* (2025).
[53] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* (2023).
[54] 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. *CoRR abs/2402.10260* (2024).
[55] Yifan Sun, Yuhang Li, Yue Zhang, Yuchen Jin, and Huan Zhang. 2024. SVIP: Towards Verifiable Inference of Open-source Large Language Models. *CoRR abs/2410.22307* (2024).
[56] Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *CoRR abs/2507.06261* (2025).
[57] Qwen Team. 2025. Qwen3 Technical Report. *CoRR abs/2505.09388* (2025).
[58] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How Does LLM Safety Training Fail?. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 80079–80110.
[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.
[60] 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* (2025).
[61] Zhikang Xie, Weilin Wan, Peizhu Gong, Weizhong Zhang, and Cheng Jin. 2026. Advanced Black-Box Tuning of Large Language Models with Limited API Calls. In *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 34079–34087.
[62] Xianjun Yang, Wei Cheng, Yue Wu, Linda Ruth Petzold, William Yang Wang, and Haifeng Chen. 2024. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. In *International Conference on Learning Representations (ICLR)*.
[63] Yage Zhang. 2025. Cutting the Root of Hallucination: Structural Trimming for Vulnerability Mitigation in Code LLMs. In *Conference on Language Modeling (COLM)*.
[64] Yage Zhang, Yukun Jiang, Zeyuan Chen, Michael Backes, Xinyue Shen, and Yang Zhang. 2026. Real Money, Fake Models: Deceptive Model Claims in Shadow APIs. *CoRR abs/2603.01919* (2026).
[65] 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, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *CoRR abs/2303.18223* (2023).
[66] 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* (2023).

A Open Science

The experiments in this paper are based on publicly available artifacts. For utility and safety evaluation, we use established benchmarks including AIME 2025, GPQA, MedQA, LegalBench, JailbreakBench, and AdvBench [10, 24, 31, 39, 51, 66], as described in Section 4. For model verification, we use two publicly available verification frameworks, LLMmap [48] and MET [30]. Non-anonymized provider evidence will be released under gated access within ethical and legal constraints. The evaluation code is available at https://github.com/TrustAIRLab/ShadowAPI.

B Ethical Considerations

In this work, we adhere to responsible AI research guidelines. Our study involves purchasing and querying unofficial shadow API services, which may violate the terms of service of official providers (e.g., OpenAI, Google, DeepSeek). These interactions are 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 limit 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.

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 risks within the community. Retrospectively identifying published papers as shadow API users carries reputational risk, so we report the 187-paper corpus only in aggregate, withholding titles, authors, repositories, and exact citation counts, and we draw no inference of misconduct or intent: when most of these papers were written, these risks were not publicly known and no norm required endpoint verification. Such findings should be read as evidence of a missing community safeguard rather than individual fault, and future audits of published research should adopt the same aggregate-only standard.

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 adhere 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.