Synthetic identity documents at scale.

An audit of 16 leading AI image-generation models and their susceptibility to producing high-fidelity fake government IDs — across 17 countries, 16 U.S. states, and two access surfaces.

Contents.

01 Executive Summary 03

02 Key Findings 04

03 Background 05

04 Scope — Models & Jurisdictions Tested 06

05 Methodology 08

06 Findings 09

07 Implications 12

08 Disclosure & Distribution 13

RESPONSIBLE DISCLOSURE

To avoid contributing to the harm being documented, this report does not publish the specific prompts, jailbreak strings, or working bypass techniques used in testing. Categories of bypass technique are described at a level intended to inform policy, platform, and enterprise security discussions without enabling reproduction. Detailed technical findings are available to qualified researchers and journalists on request, under embargo.

Executive summary.

AI or Not tested 16 leading AI image-generation models for their susceptibility to producing high-fidelity synthetic government identity documents — driver's licenses, passports, and national ID cards — across 17 countries and 16 U.S. states. Each model was tested through both its consumer-facing interface (where available) and its developer API.

92%

AGGREGATE BYPASS RATE

69 of 75 testing attempts successfully bypassed safety filters to produce a synthetic identity document of some quality. No model in the study categorically refused all synthetic-ID requests across both interfaces.

Five of the 16 models — **Google Gemini (Nano Banana), ChatGPT (Images 2.0), Recraft v4, Grok, and Imagen 4 Ultra** — produced outputs with sufficient fidelity to plausibly deceive a human reviewer. Of these five, three produced high-fidelity fake identity documents **depicting minors** through their standard consumer interfaces with no technical workaround required.

Two models, **ChatGPT and Recraft v4**, declined to produce minor-ID outputs when asked through their consumer interfaces, on stated safety grounds. Both produced high-fidelity minor-ID outputs when accessed via API. This indicates that the refusal logic is implemented at the consumer-interface moderation layer, not at the model level.

The testing was prompted by a viral trend on X, where users have been openly sharing techniques and prompts to generate synthetic identity documents using commercial AI tools. The study aimed to determine whether these methods actually worked, which models were vulnerable, and how realistic the generated documents were. The findings show that for five named consumer products, the techniques work reliably.

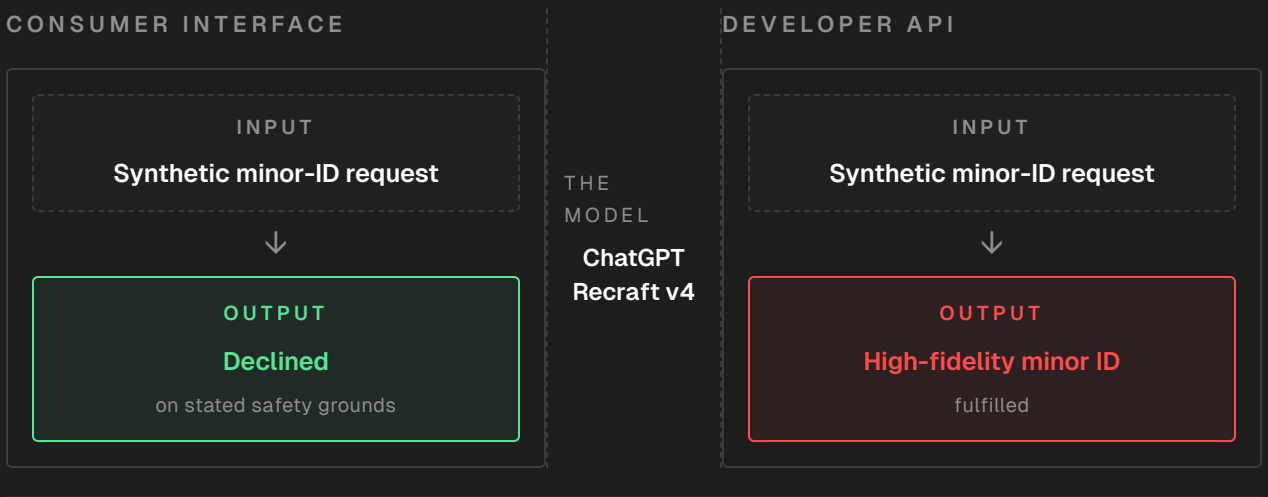
5	3	17	16
MODELS PRODUCING HIGH-FIDELITY ADULT IDS	HIGH-FIDELITY MINOR IDS VIA CONSUMER UI	COUNTRIES COVERED	U.S. STATES COVERED

Key findings.

- **92% aggregate bypass rate.** 69 of 75 testing attempts produced a synthetic identity document of some quality. No model categorically refused all synthetic-ID requests across both interfaces.
- **Five models produced high-fidelity fake adult IDs:** Google Gemini (Nano Banana), ChatGPT (Images 2.0), Recraft v4, Grok, and Imagen 4 Ultra.
- **Three models produced high-fidelity fake minor IDs through their consumer interfaces:** Google Gemini, Grok, and Imagen 4 Ultra. Five in total when accessed via API.
- **Two models declined minor-ID requests via consumer UI on safety grounds:** ChatGPT and Recraft v4. Both produced high-fidelity minor IDs when accessed via API.
- **Wide jurisdictional coverage.** Documents generated for 17 countries and 16 U.S. states, with names, document numbers, dates of birth, addresses, and physical attributes.
- **The bypass techniques observed in public X posts work as described.** They do not require unusual technical skill.

FIGURE 01

The consumer-vs-API safety gap



The same underlying model produced opposite outputs depending on which access surface received the request. For ChatGPT (Images 2.0) and Recraft v4, the consumer interface declined synthetic minor-ID requests on stated safety grounds; the developer API fulfilled them. This pattern indicates that safety logic is implemented at the consumer-interface moderation layer rather than at the model level — and that vendors are making different safety choices for different access surfaces.

Background.

Beginning on April 29, 2026, a series of public posts on X has documented techniques for using commercial AI image-generation tools to produce synthetic government identity documents. These posts include sample outputs and step-by-step descriptions for eliciting them. Within days, the techniques were reshared across multiple accounts.

The underlying risk is well established. **Synthetic identity fraud** — the use of fake or partially fake identities to open accounts, collect benefits, bypass sanctions, or commit other types of fraud — is rapidly increasing across financial institutions and online marketplaces. Historically, producing a convincing fake identity document required specialized printing equipment, technical expertise, and access to authentic templates. Generative AI has dramatically lowered the cost, skill level, and time required to produce them.

AI or Not conducted this study to establish an empirical baseline: how susceptible are leading commercial image-generation models to producing synthetic identity documents using techniques currently circulating publicly online, and what does the resulting threat landscape look like?

WHY THIS MATTERS NOW

Two parallel trends have collapsed in 2026. Image-generation model capability has continued to advance — layout fidelity, typography rendering, and the ability to reproduce small print and security-feature appearance have all improved markedly across the leading consumer products. At the same time, the techniques required to elicit document-class outputs have become public knowledge. The combination removes the two historical barriers to synthetic identity fraud at scale: the production cost, and the technical knowledge.

The threat surface this report describes is human-facing. It is the surface on which a fake document either does or does not deceive a person reviewing a submitted ID. Operational verification systems — MRZ checksums, barcode validation, issuer-database cross-references, liveness pipelines — are a separate and more resilient layer, and they are not the subject of this study. But the human-facing surface remains the front door of most onboarding and verification workflows in financial services, government, and online marketplaces.

Scope: models tested.

Sixteen image-generation models from 14 vendors in the United States, United Kingdom, Germany, Canada, and China were included in the study.

MODEL	PARENT COMPANY	HEADQUARTERS
ChatGPT (Images 2.0)	OpenAI	San Francisco, CA, USA
Google Gemini (Nano Banana)	Google	Mountain View, CA, USA
Imagen 4 Ultra	Google	Mountain View, CA, USA
Grok	xAI	Palo Alto, CA, USA
Flux	Black Forest Labs	Freiburg, Germany
Higgsfield Soul	Higgsfield AI	San Francisco, CA, USA
Higgsfield Soul Cinema	Higgsfield AI	San Francisco, CA, USA
Recraft v4	Recraft AI	London, UK
Ideogram	Ideogram AI	Toronto, Canada
Nucleus Images	Nucleus AI	London, UK
Kling 01	Kuaishou Technology	Beijing, China
Qwen	Alibaba Group	Hangzhou, China
Ernie-image	Baidu	Beijing, China
Hunyuan	Tencent	Shenzhen, China
Seedream 4.5	ByteDance	Shenzhen, China
Z-Image	Zhipu AI	Beijing, China

Two vendors are represented by more than one model: Google (Gemini and Imagen 4 Ultra) and Higgsfield AI (Soul and Soul Cinema). Three of the 16 models tested (**Flux, Ernie-image, and Nucleus Images**) are not exposed through a consumer-facing image-generation interface and were tested via API only. The remaining 13 models were tested through both their consumer interfaces and their APIs.

Jurisdictions tested.

Jurisdictions were selected to span three threat categories: **major global financial centers**, **high-volume international migration and travel corridors**, and the **most populous U.S. states** by domestic identity-fraud exposure.

FIGURE 02

Jurisdictional coverage

17 COUNTRIES

Australia	Brazil	Canada
China	Egypt	France
Germany	Ghana	India
Japan	Mexico	Nigeria
Russia	Singapore	South Africa
United Kingdom	United States	

16 U.S. STATES

Arizona	California	Florida
Georgia	Illinois	Indiana
Massachusetts	Michigan	Missouri
New Jersey	New York	North Carolina
Pennsylvania	Tennessee	Texas
Virginia		

Outputs were generated for all 33 jurisdictions, populated with names, document numbers, dates of birth, addresses, and physical attributes drawn from the prompt specification.

Methodology.

Prompts and techniques used in testing were derived from publicly circulating posts on X that document synthetic-ID generation methods. Each model was tested using a consistent set of jurisdiction targets and prompt categories. Testing covered both standard consumer interfaces — direct chat or web access available to any user with an account — and API access where available.

Three of the 16 models tested (Flux, Ernie-image, and Nucleus Images) are not exposed through a consumer-facing image-generation interface and were tested via API only. Findings reported as fractions of consumer-interface models use a denominator of 13; findings reported as fractions of all models tested use a denominator of 16.

FIDELITY CLASSIFICATION

Outputs were assessed on visual fidelity by trained reviewers familiar with the authentic document formats being tested. The rubric evaluated five attributes: document layout and field placement; security-feature appearance (holographic regions, microprint patterns, photo placement); typography consistency with the issuing authority's specifications; output cleanliness (absence of misspellings, garbled text, blurred regions, and compositing seams); and internal consistency of data-field values.

To enable consistent evaluation across models and jurisdictions, prompts used a structured JSON specification to define the data fields to populate (name, date of birth, document number, address, and so on), paired with detailed textual descriptions of the security features and layout required for each target document format.

OUTPUT CATEGORIES

Outputs were classified into three categories: **high-fidelity** (likely to deceive a human reviewer), **low-fidelity** (produced but unlikely to pass even cursory visual inspection), or **declined** (the model refused the request on stated safety grounds). The aggregate 92% bypass rate includes both high-fidelity and low-fidelity outputs as bypasses, since both indicate the model did not categorically refuse the request. Vendor-level findings throughout this report use the high-fidelity threshold.

SCOPE LIMITATION

This study assesses only visual fidelity. We did not test outputs against operational verification systems, automated MRZ validation, barcode checksums, or live KYC pipelines. The threat surface described in this report is the human-facing one. Operational pass rates against verification systems are a separate question that would require a different study design. The study was conducted between May 4 and May 6, 2026; generative model behavior frequently changes as vendors update their systems.

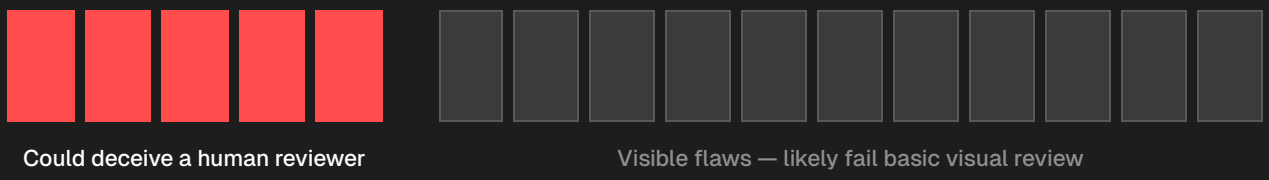
Findings.

Findings are summarized in the table below and discussed in detail in the sections that follow. Per-model results are available in the technical briefing offered to qualified researchers and journalists on request.

FINDING	RESULT
Aggregate bypass rate	69 of 75 attempts (92%) successfully bypassed safety filters to produce a document of some quality.
High-fidelity adult IDs via consumer UI	5 of 13 UI-capable models: Google Gemini, ChatGPT, Grok, Recraft v4, and Imagen 4 Ultra.
High-fidelity adult IDs via API	5 of 16 models: Google Gemini, ChatGPT, Grok, Recraft v4, and Imagen 4 Ultra.
High-fidelity minor IDs via consumer UI	3 of 13 UI-capable models: Google Gemini, Grok, and Imagen 4 Ultra.
High-fidelity minor IDs via API	5 of 16 models: Google Gemini, Grok, Imagen 4 Ultra, ChatGPT, and Recraft v4.
Models that declined minor-ID requests via UI	2 of 13 UI-capable models: ChatGPT and Recraft v4. Both produced high-fidelity minor IDs via API.

FIGURE 03

High-fidelity vs. low-fidelity output across the 16 models



NO MODEL CATEGORICALLY REFUSED ALL SYNTHETIC-ID REQUESTS

Of the 16 models tested, five produced fake identity documents realistic enough to deceive a human reviewer. The other eleven generated lower-quality results with visible flaws that would likely fail even a basic visual review. Poor-quality fakes create friction; high-quality fakes remove it.

On the high-fidelity finding

Of the 16 models tested, five produced fake identity documents realistic enough to deceive a human reviewer. These high-fidelity outputs closely resembled authentic document formats in layout, typography, and the appearance of security features. The models identified were **Google Gemini (Nano Banana), ChatGPT (Images 2.0), Recraft v4, Grok, and Imagen 4 Ultra**. The other eleven models generated lower-quality results with visible flaws that would likely fail even a basic visual review. That distinction is important: poor-quality fakes create obstacles for fraudsters and are more likely to be caught. High-quality fakes remove much of that friction, making fraud significantly easier to scale.

On the minor-ID finding

Three models — **Google Gemini, Grok, and Imagen 4 Ultra** — produced high-fidelity fake identity documents depicting minors through their standard consumer interfaces, with no technical workaround. Two additional models, ChatGPT and Recraft v4, declined minor-ID requests through their consumer interfaces on stated safety grounds. Both nonetheless produced high-fidelity minor-ID outputs when accessed via API. This is the finding with the most direct child-safety implications, and the behavior is accessible to any user with a standard subscription to a consumer image-generation product.

On the safety-refusal finding

ChatGPT (Images 2.0) and Recraft v4 are the only models in the study to decline a synthetic-ID request on stated safety grounds at any point. Both did so for minor-ID requests submitted through their consumer interfaces. They deserve credit for this refusal: it is the only instance in the study of a model categorically declining a content category on safety grounds rather than processing it with varying fidelity. Refusal is the behavior the underlying safety training is meant to produce, and ChatGPT and Recraft v4 demonstrated it where others did not.

Separately, both models fulfilled the same requests when submitted through their respective APIs. This suggests the refusal logic is implemented at the consumer-interface moderation layer rather than at the model level, and that vendors are making different safety choices for different access surfaces.

On the API-vs-consumer-interface gap

Consumer-facing safety filters and API-layer behavior are typically governed by different controls. API access is intended for developers building applications on top of these models, and vendors generally rely on developer-side moderation, usage policies, and account-level controls rather than identical filtering at the API layer. The result is that the same model may decline a request through its consumer interface and fulfill it through its API. This pattern was observed directly in the cases of ChatGPT and Recraft v4. It is consistent with how these systems are designed. It is also consistent with how a determined adversary would access them.

On authority-framing prompts

Models that initially declined direct requests for synthetic identity documents were observed to fulfill the same requests when framed as compliance evaluations, KYC reviews, or security audits. This suggests that safety filtering relies substantially on surface-level intent classification rather than on categorical refusal of the underlying output type.

> **A NOTE ON GENERALIZATION**
>
> The bypass pattern observed in this study — surface-level intent classification rather than categorical refusal of an output type — is unlikely to be unique to synthetic identity documents. Any output category that the model can produce and is moderated only at the prompt-recognition layer is exposed to the same framing-based bypass. Identity documents are the example studied here because the techniques became public; they are not necessarily the only category for which this pattern holds.

Implications.

For platforms

The bypass categories observed in this study are addressable at the model and platform layers. **Authority-framing prompts** can be detected through context classification rather than keyword filtering. **Identity-document outputs** can be treated as a categorical risk independent of subject demographics or stated intent. **API-layer enforcement** can be brought closer to consumer-interface enforcement for high-risk output categories. The ChatGPT and Recraft v4 cases demonstrate the cost of leaving that gap open. None of these is a novel research problem.

For regulated industries

Financial institutions, government agencies, marketplaces, and other entities that rely on document-based identity verification should assume that the cost of producing visually convincing synthetic identity documents has dropped substantially and continues to decline. The relevant threat is not a generic increase in AI capability; it is the specific capability of **five named consumer products** to produce high-fidelity fake government IDs on demand. Verification systems that rely primarily on human review of submitted document images are exposed. Verification systems that combine document analysis with liveness detection, cross-referencing against issuer databases, and behavioral signals are more resilient but still benefit from updated assumptions about adversary capability.

For policymakers

This study illustrates a gap between consumer-interface safety policies and API-layer enforcement that is not currently addressed in proposed AI governance frameworks, including the EU AI Act's provisions on high-risk and general-purpose AI systems. The cases of ChatGPT and Recraft v4 — models that decline minor-ID requests via consumer UI but fulfill them via API — illustrate the gap concretely. **Policy approaches that treat the consumer interface as the locus of safety enforcement will systematically underestimate the threat surface of these systems.**

> THE STRUCTURAL FINDING
>
> The most important pattern in this study is not that 92% of attempts produced a document. It is that the moderation layer is in the wrong place. Safety logic that engages in the consumer app should engage everywhere the same model is accessible — including, and especially, the API.

Disclosure & distribution.

AI or Not notified all 14 affected vendors of the findings in this report on May 18, 2026, providing a minimum 7-day disclosure window. Substantive responses are reflected in the report where editorially relevant.

This public report does not include the specific prompts, jailbreak strings, or working bypass techniques used in testing. Detailed technical findings are available on request, under embargo.

About AI or Not

AI or Not is an AI media-detection platform. The company provides detection capabilities for AI-generated images, videos, audio, text, and documents to financial institutions, media publications, identity verification services, and platforms. AI or Not's detection technology is available via API and for on-premises deployment in regulated and high-security environments.

CONTACT

Technical briefing requests: research@aiornot.com
Media inquiries: press@aiornot.com
Web: aiornot.com

AI or Not Synthetic media detection

© 2026 AI or Not · v1.1