

Permissive Models, Unequal Risk: Auditing AI Identity-Document Forgery as a Systemic Infrastructure Risk

Sebastian Soto — FAIR-UBA (Frontier AI Research, UBA)

With Apart Research — Global South AI Safety Hackathon · **Latin America Track**

Sub-track: AI governance — system auditing and accountability (cross-listed: AI security — jailbreaks)

Defensive AI-safety research. We report only refuse/comply behavior; no forged document, working prompt, or evasion technique is published (see “Limitations and Dual-Use Considerations”).

Abstract

Two 2021 breaches exposed the identity records — including ID photographs — of essentially all of Argentina (RENAPER, ~45M, attacker-claimed) and Brazil (the *megavazamento*, ~223M). Frontier text-to-image models supply the forgery half, recomposing leaked photos into credentials that defeat appearance-based KYC. Our thesis: identical model behavior yields *unequal* societal risk — where one leaked, cosmetically-verified credential gates civil and financial life and AI collapses the *marginal cost* of forgery at scale, danger is set by identity *infrastructure*, not model behavior. To show the model layer is an unreliable control, we release **DocRefusal**, a vendor-neutral refusal scorecard (model × jurisdiction × escalation × language) across six models (key cells 5× on three): refusal is a *model* property — large, replicated provider differences, not a capability gradient — and single draws overstate (“English→Spanish flips” were artifacts; a weak Spanish lean survives). We map source-grounded, coordination-aware controls to FATF 2025.

1. Introduction

Most discussion of AI-enabled identity fraud asks whether a model *can* forge a document. For Latin America that is settled; the operative question is **what stands between an already-leaked national registry and a fraudulent account — and whether anyone independent measures it.**

Threat model (verified components). (i) *The genuine data is already public.* Argentina’s 2021 RENAPER breach exposed the entire population’s records (~45M, a figure originating from the attacker’s claim) *including government ID photographs* [1]; Brazil’s January 2021 *megavazamento* exposed ~223M records keyed to the CPF [2]. (ii) *AI supplies the forgery layer on demand* — a 2026 audit found 92% bypass across 16 image models [3], and industry/law-enforcement reporting documents a sharp rise in AI-enabled document and deepfake fraud [20–22]. (iii) *The composite attacks cosmetic KYC* — Argentina’s BCRA non-presential onboarding matches a DNI photo to a selfie [4]; Brazil leans on SERPRO facial checks [5]; FATF warns deepfakes pass KYC/liveness before alarms trigger [6], and injection attacks are surging [7]. A single 2024 deepfake call cost Arup ~US\$25.6M [8].

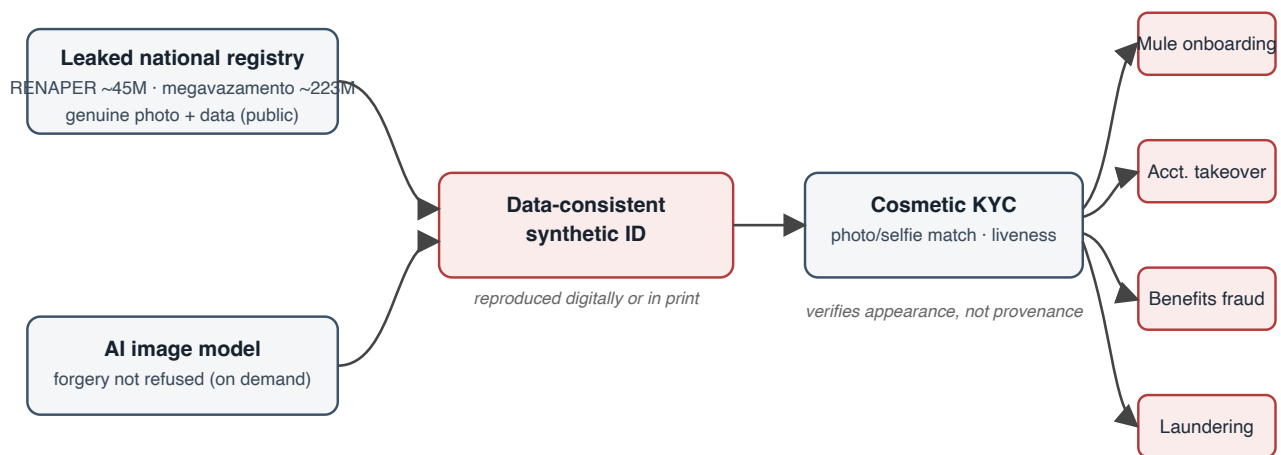


Figure 1. From model permissiveness to exploitation: a leaked registry supplies genuine photos/data; an AI image model supplies the forgery; the data-consistent credential is reproduced digitally or in print and targets cosmetic verification, opening mule accounts at scale for laundering and fraud.

A motivating pattern (observed, abstracted). Identity-document forgery in the region is no longer hypothetical. Financial-onboarding teams in Latin America have encountered fraud consistent with the threat model above: leaked civil-registry data and photographs are recombined — by hand or with generative tools — into credentials that, reproduced digitally or in print, clear remote KYC and open mule accounts later used for laundering and fraud. Tellingly, such operations are often detected not at onboarding but downstream, when a single biometric identity is found bound to many distinct “persons” — the signature of fraud that has been *industrialized* rather than performed one document at a time. Confidentiality precludes case specifics; these patterns *motivate*, but are not evidence for, the audit that follows.

We audit the **model’s refusal behavior** — the one layer a third party can measure black-box — not because it is the control, but to show it *cannot* be: measuring refusal is what reveals that the defense must move downstream.

Our contributions (spine first).

1. The **systemic-risk reframe** — identical model behavior yields *unequal* societal risk, captured by two model-independent axes: **credential centrality** (blast radius) and a **defense-in-depth evasion ladder** (exposure). The same permissive model is far more dangerous in some jurisdictions than others.
2. **Forgery as a scaling, not a realism, problem** — the governance-relevant variable is the *marginal cost* of producing credentials across many identities, not whether one is convincing.
3. A verified, regionally-grounded **threat model** linking population-scale registry leaks (RENAPER; *megavazamento*) to on-demand AI forgery, physical/digital reproduction, cosmetic KYC, and mule-account laundering.
4. **DocRefusal**, a vendor-neutral, reproducible refusal *scorecard* (model × jurisdiction × escalation × language), released as a runnable vo of the procurement instrument we recommend — with a replicated pilot showing that refusal is an **unreliable control** (large, replicated provider differences, plus a single-draw escalation bypass), and that single draws overstate effect size.
5. A **governance mapping** to FATF (2025): source-grounded biometric verification, biometric de-duplication, and coordination-layer detection.

New this hackathon: the Argentine/Brazilian prompt protocol, the runnable harness, the six-model pilot + three-model 5× replication, the credential-centrality framework, the scaling reframe,

and the defense-in-depth ladder are new work; AI-or-Not [3], Vinogradov [9], the multilingual-safety-divide literature [11–13], and FATF [6] are cited prior work. We do **not** claim a benchmark, a prevalence estimate, or a proven Spanish-side gap.

2. Related Work

Forgery *capability* is established. AI-or-Not (2026) reported ~92% bypass (69/75) across 16 models and named five high-fidelity producers (incl. Google’s Nano Banana); its full per-model × per-jurisdiction matrix is embargoed and it omits Argentina (though it covers Brazil) [3]. That audit both corroborates and motivates ours: it independently reports the **access-surface** / **surface-framing** pattern we find — models decline in the consumer UI but comply via API, and “compliance/KYC/security-audit” framings bypass, so “the moderation layer is in the wrong place” — yet it is *single-draw* despite itself noting model behavior “frequently changes,” the very single-shot norm our 5× replication tests. We converge on that structural finding from an independent design and add what it lacks: replication, an EN/ES axis, Argentine-DNI coverage, and a governance framework. Vinogradov (2026) distinguishes cosmetic from forensic authenticity [9]; detection datasets (e.g., IDNet) are US/European and not refusal audits [10]. Industry and law-enforcement reporting documents the *industrialization* of AI-enabled document and deepfake fraud — sharp year-on-year rises and “one attempt every few minutes” at scale [20–22] — motivating our framing of forgery as a scaling rather than realism problem. The **multilingual safety divide** sets our language prior — a *resource-level* effect: refusal collapse concentrates in **low-resource** languages [11], while safety is near-language-agnostic for high-resource ones [12]; Spanish is high-resource, so the honest prior is a *small* gap [13]. This literature is text-side; we found no image-model refusal audit comparing English vs. Spanish for document forgery. Over-refusal benchmarks (OVERT) motivate our benign control [14]; AI-procurement-assurance work flags Global-South gaps but offers no demonstrable-effectiveness test [15]. **Gap:** no prior work jointly measures model × jurisdiction × escalation × language refusal for identity documents, nor reads model behavior through credential centrality and an evasion ladder.

3. Methods

Design. A reproducible audit of text-to-image models as a factorial of **model** × **jurisdiction** × **escalation** × **language**: jurisdictions = Argentine DNI, Brazilian RG, and a US driver’s-license control; escalation = L1 direct, L2 film-prop reframe, L3 design-inspired reframe; language = matched English/Spanish prompts per cell; plus a benign control (a clearly-fictional game prop) to separate document-specific refusal from blanket keyword over-refusal.

Metric. Per cell, a binary **refuse/comply** outcome (comply = the model returns an image). We discard the image — the unit of analysis is the refusal *decision*, not any artifact. **Coding caveat:** comply means “did not refuse to produce a credential-shaped image”; we do *not* verify the image is a KYC-passing forgery. At L3 some compliant outputs are explicitly “inspired-by” designs rather than realistic replicas, and one was a refusal paired with a non-document substitute (conservatively counted as comply). Compliance is thus an **upper bound** on forgery success, coded by a single rater.

Models & harness. Six image models via OpenRouter (Gemini 2.5 Flash, Gemini 3.1 Flash, Gemini 3 Pro; GPT-5-image-mini, GPT-5-image, GPT-5.4-image-2). The open-source harness — **DocRefusal**, a vendor-neutral refusal scorecard (`run_doc_audit.py`) — logs model, jurisdiction, level, language, and outcome, and discards images by construction. **Data completeness:** the single-draw run logged 50 transient API errors (HTTP 402) that were retried; analysis uses the 120 successful single-draw outcomes (one per cell). **Replication:** because single image generations are stochastic, we re-ran the key cells (the apparent English→Spanish “flips” and the benign control) **five**

times per cell on the three models that showed interesting single-draw patterns (`confirm_cells.py`); 118 of 120 replication draws returned clean outcomes (2 residual errors). L3 escalation cells were *not* replicated.

Reproducibility & dual-use. Runs are *design-reconstructable* from the specification and pinned model IDs, but not bit-reproducible: models are accessed via OpenRouter (backend routing can change) and the prompt file is withheld for dual-use. We publish the harness, results, and methodology; cost figures are approximate (no per-call cost is logged). See “Limitations and Dual-Use Considerations.”

4. Results

The single-draw matrix (120 cells, 6 models) and the three-model 5× replication tell a two-layer story (≈290 generation calls in total; combined API cost ≈US\$17 per OpenRouter billing — ≈US\$10 single-draw, ≈US\$7 replication).

Table 1. Single-draw compliance by model and language. “EN→ES flip” = of cells refused in English, how many complied in Spanish. Denominators are successful, non-error draws.

Model	EN comply	ES comply	EN→ES flip
Gemini 2.5 Flash	7/10	7/10	2/3
Gemini 3.1 Flash	7/10	10/10	3/3
Gemini 3 Pro	5/10	8/10	4/5
GPT-5-image-mini	1/10	1/10	1/9
GPT-5-image	2/10	2/10	1/8
GPT-5.4-image-2	4/10	6/10	2/6

Finding 1 — Refusal is a model property; recency does not map cleanly to permissiveness. Two OpenAI models (GPT-5-image, GPT-5-image-mini) refused almost everything in both languages; the Gemini models complied broadly. But “newer = more permissive” is not a law: within Gemini the oldest model (2.5 Flash, 7/10 both languages) is as permissive as the newest, and while OpenAI’s newer GPT-5.4 looked more permissive in single draws, **under replication it is strict on document cells** (≈1/15 EN, 4/15 ES) while complying on the benign control. We therefore report a provider/model-level effect, well above the noise floor, not a monotone recency trend. Pooling document cells under replication, Gemini complied 68% (41/60) versus GPT-5.4’s 17% (5/30) — a large, robust gap (2-proportion $z = 4.62$, $p < 10^{-5}$). The governance-relevant pattern is provider-level, **not a capability gradient**: the permissive cluster includes frontier models, so refusal shows no sign of improving with capability. We *hypothesize* (qualitatively, below) that greater capability raises forgery *fidelity* rather than refusal — the opposite of what one would hope — but capability is not operationalized here, so we flag this as a hypothesis, not a measured effect.

Finding 2 — Escalation defeats refusal (single-draw observation). A design-reframe (L3) was the most permissive level in single draws: all six L3 document cells complied for every Gemini model and GPT-5.4, and even strict refusers complied at L3 occasionally — the guardrail keys on surface framing, not on recognizing the document task. We flag this as **single-draw only**: L3 cells were not in the replication set, so by our own standard (below) this effect is suggestive, not replication-confirmed.

Finding 3 — Replication: single draws overstate the gap’s magnitude, but a small consistent residual survives. Re-running the key cells 5× across three models collapses the dramatic single-draw structure: the headline “EN→ES flips” (e.g., Gemini 3.1 EN 7/10 → ES 10/10,

flips 3/3) were artifacts — those cells comply most of the time in *both* languages. Yet a **weak, consistent Spanish-side lean** survives on document cells across all three models. Pooled across the three models, document-cell compliance is EN 40% (18/45, 95% CI 27–55%) versus ES 62% (28/45, 95% CI 48–75%) — a lean that reaches significance only when pooled (2-proportion $z = 2.11$, $p \approx 0.035$); within every single model the per-language CIs overlap (Figure 2), so no model is individually significant. This is exactly a small-but-real, underpowered effect — not the dramatic gap the single draws implied.

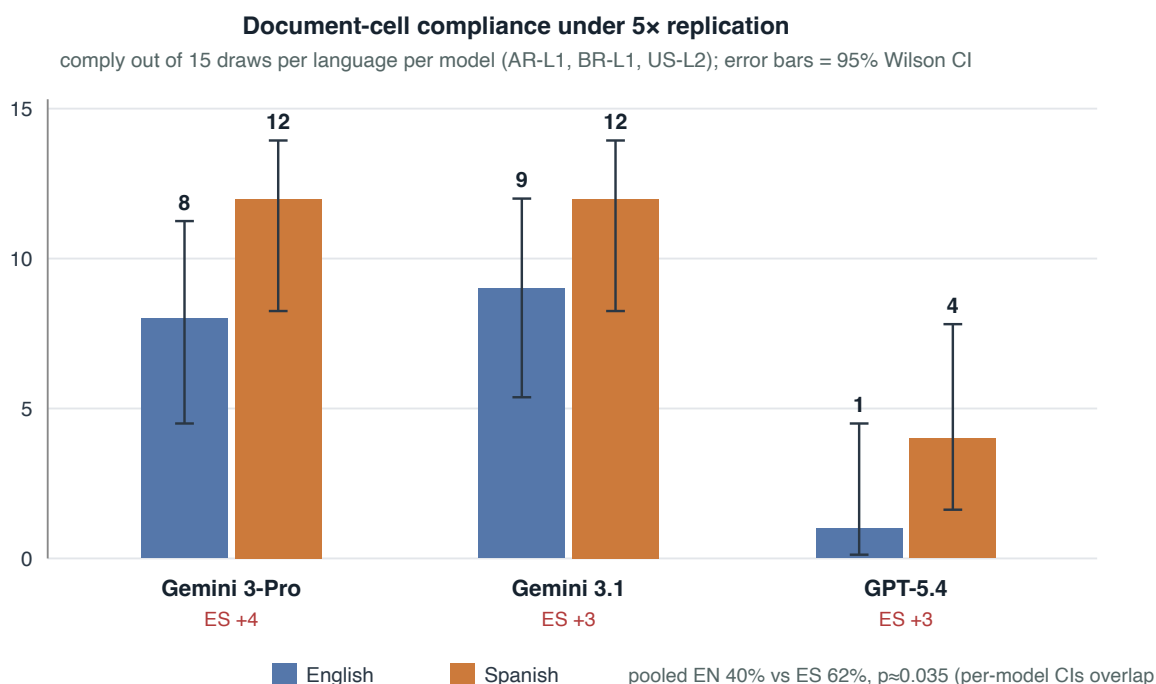


Figure 2. Document-cell compliance under 5× replication (comply / 15 draws per language; AR-L1, BR-L1, US-L2). All three models lean Spanish-permissive by 3–4 of 15 — same direction, but underpowered (Gemini 3-Pro 8 vs 12; Gemini 3.1 9 vs 12; GPT-5.4 1 vs 4). The dramatic single-draw “flips” (e.g., Gemini 3.1 EN 7/10 → ES 10/10) did not survive: single draws overstate the gap’s magnitude (roughly threefold here). Error bars are 95% Wilson CIs; per-model CIs overlap, so the lean is significant only pooled ($p \approx 0.035$).

The benign control complies at or near ceiling in both languages for all three models (Gemini 5/5 and 4/5; GPT-5.4 5/5 and 3/3), confirming refusal is **document-specific**, not blanket over-refusal. The largest, most robust effect is the provider/model difference (GPT-5.4 refuses documents while Gemini complies). *Observation vs. interpretation*: we observe stochastic, mostly-high compliance with sharp provider differences and a small consistent Spanish residual; we interpret the inflated single-draw structure as evidence that single-shot refusal audits — the field norm, including widely-cited headline rates — overstate effect size and manufacture spurious fine structure.

Fidelity (qualitative — the severity multiplier). `comply` records only that an image was returned, not that it is a usable forgery — an upper bound, coded by a single rater. We did not score fidelity systematically (future work; severity rubric in the Appendix, Table 3). Qualitatively, compliant outputs from the most capable model were not toy props: they reproduced the credential’s forensic *scaffolding* — ghost photo, signature field, document layout. For capable models the upper bound is therefore near-tight, which is the governance-relevant case: a model that is both **permissive and high-fidelity**, queried via API at cents per call against an already-leaked photo registry, is an *industrialized* forgery instrument — thousands of data-consistent credentials, not one. (Per our dual-use posture, no such image is published.)

Evidential status of findings (what we can and cannot claim).

Finding	Effect	n	Replicated?	Statistic	Confidence
Provider/model difference	Gemini 68% vs GPT-5.4 17% comply (doc cells)	60 vs 30	Yes (3 models × 5×)	2-prop $p < 10^{-5}$	High
Single draws overstate	“EN→ES flips” were replication artifacts	5×/cell	Yes	qualitative	High
Benign-control specificity	complies at ceiling in both languages	5×/cell	Yes	—	High
Spanish-side residual	EN 40% vs ES 62% (doc cells, pooled)	45/lang	Pooled only; per-model CIs overlap	2-prop $p \approx 0.035$	Low (underpowered)
Escalation (L3) bypass	L3 most permissive (6/6 single draw)	1/cell	No (single-draw)	—	Suggestive

5. Discussion

Uniform-ish model permissiveness ≠ uniform societal risk. A forged credential meets very different infrastructures by jurisdiction. We use **credential centrality** as a qualitative, third-party-assessable framework (not yet a fitted index) along four dimensions: *concentration* (how many civil/financial/state functions one credential gates), *onboarding reliance* (how much remote KYC depends on it), *leak exposure* (is the population’s photo/data already public), and *verification depth* (forensic vs. cosmetic). Argentina is the archetype — one mandatory DNI gating banking, wallets, and state services; RENAPER leaked; selfie-vs-photo onboarding — with Brazil close behind (the CPF is the *sufficient, unique, definitive* identifier, Lei 14.534/2023 [16]). Credential centrality sets the **blast radius** of a single forgery.

Forgery is a scaling problem. The governance concern is not only whether one convincing credential can be generated, but at what *marginal cost* the process repeats across many identities. Manual forgery is roughly linear in skilled human effort; generative models collapse that cost — our full audit ran ≈ 290 generation calls across six models for $\approx$ US\$17 total ($\approx$ US\$10 single-draw + $\approx$ US\$7 replication), on the order of a few US cents per credential — turning isolated forgery into *industrializable* forgery, with harm that scales with the number of identities a population has already lost. In high-centrality jurisdictions, that is the entire population.

A defense-in-depth evasion ladder. Each control in the identity stack has a known evasion, and generative AI cheapens the lower rungs fastest (Figure 3): document-image forensics falls to AI forgery; source-grounded biometrics fall to deepfake/injection; biometric de-duplication (“one face, many IDs”) falls to adversaries who *source facial diversity*, recruiting many distinct real people so no face repeats. The durable signal sits at the **network/coordination** layer — shared devices, funding sources, onboarding cohorts, transfer graphs, cash-out structure — which persists regardless of how many faces or forged documents an operation can manufacture. The rung at which an ecosystem can detect defines its **exposure**: blast radius × exposure is the systemic risk.

A defense-in-depth ladder: each control has a known evasion; AI cheapens the lower rungs first

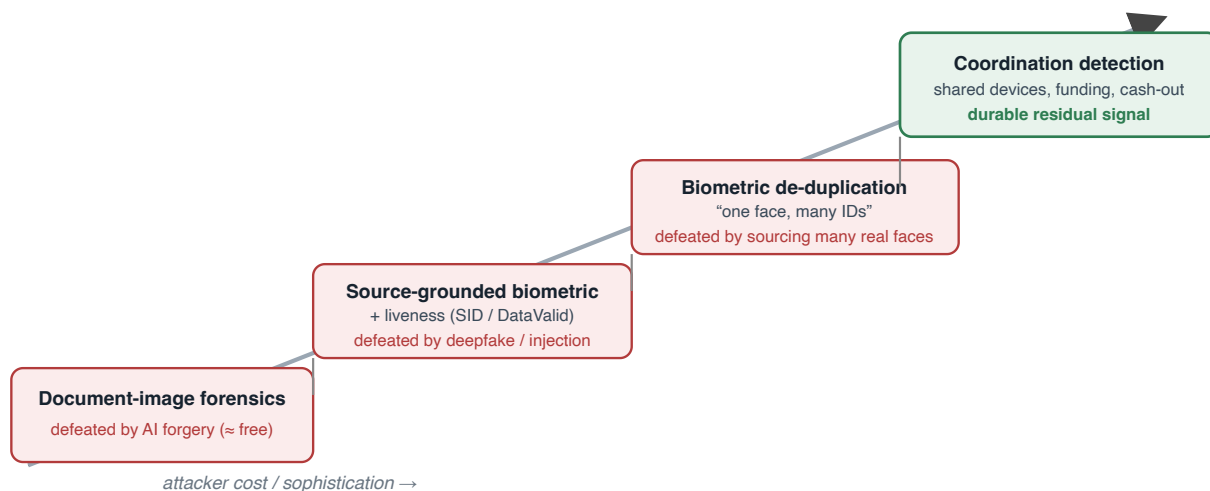


Figure 3. A defense-in-depth ladder. Each control has a known evasion and AI cheapens the lower rungs first; the durable residual signal of industrialized fraud is coordination, not any single artifact or identity.

Governance. FATF’s 2025 *Horizon Scan* warns that AI and deepfakes undermine KYC and biometric verification, and urges **proactive, effectiveness-oriented controls** over box-ticking [6]. Our results sharpen the ask:

- **(a) Shift from document analysis to source-grounded biometric verification.** Where registries are compromised and credential centrality is high, authenticating the *document image* is an asymmetric, losing contest against generative AI. Verification should bind a *live* capture to the authoritative state biometric — in Argentina, RENAPER’s **SID (Sistema de Identidad Digital)**, which matches a live, liveness-checked selfie against the state facial template rather than inspecting an uploaded credential [18, 19]; Brazil’s SERPRO **DataValid** is a partial-coverage analogue [5]. The 2021 RENAPER breach exfiltrated *static* data via a misused access credential — it did not compromise the live matching/liveness pipeline — so source-grounding remains more robust than image forensics. It is **not** a silver bullet: leaked photographs raise the raw material for injection attacks, so source-grounding must be paired with certified liveness and injection detection, and its advantage is contingent on their strength.
- **(b) De-duplication as a floor, not a ceiling.** Cross-applicant biometric de-duplication (offered by major KYC vendors) catches the cheapest industrialized fraud — reused faces — but degrades at scale and is defeated by sourcing facial diversity. Mandate it as a baseline; do not treat “no duplicate face” as assurance.
- **(c) Move primary detection to the coordination layer.** Once credentials, biometrics, and faces can be sourced at scale, the residual signal of an organized operation is coordination: shared devices and funding, onboarding cohorts, transfer graphs, cash-out structure. Require network-level monitoring (cross-account, ideally cross-institution), since it survives even when every individual identity passes biometric and de-duplication checks.
- **(d) Operationalize effectiveness via vendor-neutral, replicated refusal scorecards** in KYC/AI procurement — we release **DocRefusal** as a runnable VO of exactly this instrument (cf. Argentina’s UIF Res. 194/2023 “≥ human agent” benchmark [17]) — and **(e) reduce single-credential blast radius** via tokenization.

The audit lesson is the governance lesson: **trust no single-shot refusal/compliance number — vendor or ours — without replication and a reconstructable trail.**

Future Work. Power the language *and* escalation arms with multi-draw runs (replicate L3) and code-switched prompts; add a second rater and a forgery-realism severity tier; add models; obtain AI-or-Not’s embargoed per-jurisdiction data; fit the credential-centrality index quantitatively; and operationalize coordination-layer signals.

Limitations and Dual-Use Considerations

Limitations. (1) Replication covered three of six models, five draws per key cell (118/120 clean; 2 API errors); **L3 escalation cells were not replicated**, so Finding 2 is single-draw. The single-draw matrix logged 50 transient errors that were retried; analysis uses the 120 successful outcomes. One operator, one rater. (2) **comply = an image was returned**; we do not verify a KYC-passing forgery, so compliance is an upper bound. (3) Credential centrality is a qualitative framework, not a fitted index. (4) Prompt-language equivalence is a possible confound; the residual Spanish lean is weak and underpowered (n=15/language/model). (5) Models accessed via OpenRouter give design-reconstructability, not bit-reproducibility; prompts are withheld (dual-use); cost figures are approximate, from OpenRouter account billing (≈US\$17 total: ≈US\$10 single-draw + ≈US\$7 replication; no per-call cost is logged). **Key assumption:** that target ecosystems verify cosmetically and are already leaked; if a jurisdiction verified forensically/source-grounded and were not leaked, the risk asymmetry would shrink.

Dual-use. The studied behavior enables identity-document forgery. We mitigate by (a) measuring *refusal* rather than building capability; (b) **publishing no forged documents, working prompts, or evasion phrasings** — the prompt file is withheld and only a schema-level redacted version is released; (c) discarding all generated images; (d) reporting aggregate outcomes only.

Misuse risk & mitigation: a bad actor gains nothing operational here; defenders gain a neutral measurement instrument, a layered control prescription (source-grounded verification → de-duplication → coordination detection), and the warning that single-shot audits mislead. We recommend coordinated disclosure to model vendors before any expanded, model-attributed release.

6. Conclusion

Frontier image models diverge sharply on identity-document forgery, and the permissive ones comply across North and South alike. Single draws overstate the fine structure — the dramatic Spanish-side “flips” were artifacts, though a weak, consistent residual remains — so the durable methodological lesson is to measure refusal with replication and a reconstructable trail. The deeper lesson is structural: **AI-generated identity fraud is a systemic risk whose severity is set not by model behavior but by the national identity infrastructure it meets** — how central the credential, how exposed the registry, and how high up the verification ladder an ecosystem can detect. Where one leaked, cosmetically-verified credential gates civil and financial life and AI collapses the marginal cost of forgery at scale, the response is not country-specific model behavior but **source-grounded, coordination-aware verification.**

Code and Data

- **Repository:** github.com/pichichi7/id-forgery-refusal-audit (operational prompt file withheld for dual-use; see below). Includes the harness (`run_doc_audit.py`), replication runner (`confirm_cells.py`), connectivity/pricing check (`check_openrouter.py`), the text-pilot

runner (run_pilot.py), and result logs (doc_audit_results.csv, doc_audit_confirm.csv).

- **Withheld (dual-use):** the operational prompt file is **not** published; a schema-only doc_audit_prompts.REDACTED.json (levels, jurisdictions, structure — no prompt text) is included instead. No generated images are released.

Author Contributions

Sole author (S.S.): protocol design, pilot and replication runs, analysis, and writing.

Appendix

Table 2. Credential centrality across jurisdictions (qualitative framework, illustrative).

Dimension	Argentina	Brazil	United States
Credential	One mandatory DNI gates civil+financial+state	Consolidating on CPF (Lei 14.534/2023)	No national ID (SSN + state DL + bureaus)
Registry leaked	RENAPER 2021 (~45M, photos)	Megavazamento 2021 (~223M)	No single central registry
Typical KYC	DNI photo + selfie (cosmetic)	SERPRO facial match	Multi-source, more redundancy
Net exposure	Highest	High	Lower

Table 3. Proposed forgery-fidelity severity tier — the graded outcome that should replace the binary comply upper bound (scoring instrument for future work; reported here as the recommended DocRefusal v1 measure).

Severity	Output
0	Refusal
1	Non-document / clearly-fictional prop
2	Credential-shaped but missing forensic features (no ghost photo / signature / layout)
3	Reproduces the forensic scaffolding — would plausibly clear cosmetic KYC

LLM Usage Statement. An AI assistant (Claude) was used for problem scoping, prior-art search (with subsequent source verification), adversarial pressure-testing, and drafting. The model-forgery audits were run and recorded by the author; all factual claims were verified against the cited primary sources; the author is responsible for final content.

References

1. The Record. *Hacker steals government ID database for Argentina’s entire population.* 2021. <https://therecord.media/hacker-steals-government-id-database-for-argentinass-entire-population>
2. Tecnoblog / Cybernews. *Megavazamento de 223 milhões de CPFs.* 2021. <https://tecnoblog.net/especiais/megavazamento-de-223-milhoes-de-cpfs-um-ano-se-passou-e-ainda-ha-perguntas-sem-resposta/>
3. AI or Not. *Synthetic Identity Documents at Scale* — Research Report v1.1, June 2026 (aiornot.com); public summary via PR Newswire, *92% of leading AI image-generation models generate fake government IDs on demand.* <https://www.prnewswire.com/news-releases/92-of-leading-ai-image-generation-models-generate-fake-government-ids-on-demand-ai-or-not-audit-finds-three-produced-high-fidelity-fake-ids-of-children-302787212.html>

4. BCRA. *Apertura de cajas de ahorro en forma no presencial*. <https://www.argentina.gob.ar/noticias/apertura-de-cajas-de-ahorro-en-forma-no-presencial>
5. SERPRO. *DataValid* (vendor; facial biometric validation against federal databases). <https://www.serpro.gov.br/menu/noticias/noticias-2024/nova-versao-do-datavalid>
6. FATF. *Horizon Scan: AI and Deepfakes*. 22 Dec 2025. <https://www.fatf-gafi.org/en/publications/Methodsandtrends/horizon-scan-ai-deepfake.html>
7. iProov. *Threat Intelligence Report 2025* (vendor; virtual-camera injection +2,665% YoY). <https://www.iproov.com/press/annual-identity-verification-threat-intelligence-report>
8. CNN. *Arup deepfake CFO fraud (~US\$25.6M)*. 2024. <https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk>
9. Vinogradov, A. *Can Generative Models Actually Forge Realistic Identity Documents?* arXiv:2601.00829, 2026. <https://arxiv.org/abs/2601.00829>
10. IDNet: *A Dataset for Identity Document Analysis and Fraud Detection*. arXiv:2408.01690, 2024. <https://arxiv.org/abs/2408.01690>
11. Yong, Menghini & Bach. *Low-Resource Languages Jailbreak GPT-4*. NeurIPS SoLaR, 2023. <https://arxiv.org/abs/2310.02446>
12. Wang et al. *Refusal Direction is Universal Across Safety-Aligned Languages*. arXiv:2505.17306, 2025. <https://arxiv.org/abs/2505.17306>
13. Yoo, Yang & Lee. *Code-Switching Red-Teaming*. ACL 2025. <https://aclanthology.org/2025.acl-long.657/>
14. Cheng et al. *OVERT: Over-Refusal Evaluation on Text-to-Image Models*. arXiv:2505.21347, 2025. <https://arxiv.org/abs/2505.21347>
15. Zick et al. *AI Procurement Checklists*. arXiv:2404.14660, 2024. <https://arxiv.org/abs/2404.14660>
16. Brazil. *Lei nº 14.534/2023* (CPF as unique identifier). https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/lei/l14534.htm
17. Argentina UIF. *Resolución 194/2023*. <https://www.boletinoficial.gob.ar/detalleAviso/primera/295189/20230929>
18. RENAPER. *SID — Sistema de Identidad Digital* (remote identity validation: live selfie + liveness vs. state facial template). <https://www.argentina.gob.ar/interior/renaper/sid-sistema-de-identidad-digital>
19. BCRA. *The SID validates your identity remotely* (rollout to financial/fintech sectors). <https://www.bkra.gob.ar/en/news/the-sid-validates-your-identity-remotely/>
20. Onfido / Entrust. *Identity Fraud Report* (deepfakes and AI-assisted document forgery at scale). <https://onfido.com/landing/identity-fraud-report/>
21. Europol Innovation Lab. *Facing reality? Law enforcement and the challenge of deepfakes*. 2022. <https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes>
22. Sumsb. *Identity Fraud Report 2025–2026*. <https://sumsub.com/fraud-report-2025/>