

Innocent Words, Harmful Data: How Outsourced Review and Regional Blind Spots Let Coded Manipulation Slip Into AI Pipelines

Yasmine Badawy

PhD Student in Cyber Security, School of Technology & Engineering, National University

With

Apart Research

Abstract

Some of the most critical safety risks in AI development originate from training data that appears completely benign on the surface. While recent safety literature demonstrates that secret loyalties and hidden behaviors persist through post-training alignment, checking pre-training data remains the primary defense. However, current data-sanitization processes suffer from systemic structural flaws: safety review relies heavily on low-paid contract workers executing rigid legal templates via machine translation tools like Google Translate without expert leadership. This creates massive blind spots in regional, dialectal, and culturally coded language. While low-paid reviewers remain necessary for high-volume first-pass triage, they cannot be left in charge of context-sensitive safety decisions. By examining real-world pipeline failures—ranging from "wife voice" registers masking commercial sex coercion, to machine translation stripping regional slang, to academic queries about ancient royalty escalating into the modern encouragement of incest—we illustrate how clean-looking, manipulative text evades surface filters. We propose a two-tiered governance fix: (1) placing language-specific expert trios (a Historian, a Political Scientist, and a Local Linguist) in direct charge of data annotation leadership per target language, and (2) establishing independent auditing partnerships with universities and government bodies. By combining institutional authority with specialized regional expertise, this framework creates a strong deterrent that stops manipulative behavior before deployment.

1. Introduction

Some of the most dangerous training data does not look dangerous. An AI model can absorb harmful, biased, or covert behavioral traits from text that contains no offensive vocabulary,

because the harm resides entirely in the underlying context, register, and regional meaning. This vector is exemplified by the installation of narrow secret loyalties, which are formally defined as activation-conditioned distributional shifts that permanently bias model generations toward a target principal while leaving baseline assistant personas completely unaltered on benign user inputs (Lamerton & Roger, 2026).

Two recent findings establish the severity of this threat:

1. **Persistence of Hidden Objectives:** Hidden goals or "secret loyalties" trained into a model are exceptionally difficult to audit, detect, or remove post-training. A model can quietly serve a targeted principal or execute covert behaviors while maintaining standard performance during audits (Hubinger et al., 2024; Lamerton & Roger, 2026).
2. **Data Sanitization as the Primary Defense:** Because narrow secret loyalties remain completely dormant and indistinguishable from baseline models under typical black-box interrogations, evaluating and filtering dataset structures before model optimization remains the primary layer of defense (Lamerton & Roger, 2026).

The Structural Blind Spot: Both automated reward models and human-in-the-loop review (e.g., RLHF) overwhelmingly judge data based on surface-level markers—such as slurs, explicit keywords, or overt threats. This paper identifies two interconnected systemic failures that allow harmful data to slip past data checks:

- **Unguided Labor Dynamics:** Data annotation relies heavily on precarious, low-wage contract workers who lack domain expertise, operate under strict quotas, and blindly execute rigid legal guidelines without expert leadership.
- **Linguistic and Cultural Flattening:** AI developers routinely apply Western-centric safety rules to foreign-language pipelines, relying on machine translation (e.g., Google Translate) that flattens regional slang, dialectal nuance, and coded coercion into clean, benign prose.

Our Main Contributions Are:

- **An Industry Analysis of Data Annotation Blind Spots:** We detail how unguided contract labor, legal template misinterpretations, and machine-translation flattening allow toxic or manipulative data to pass surface inspections.
- **Concrete Case Studies in Coded Harm:** We demonstrate failure modes where surface-clean text masks severe harm—including "wife voice" registers concealing commercial sex trade listings, historical prompts generating bias, and conflict-zone dialect suppression.
- **An Expert-Led, Institutionally Audited Governance Framework:** We propose an operational governance model where low-cost triage is governed by **language-specific expert trios** (a Historian, a Political Scientist, and a Local Linguist assigned per target language/region) and backed by independent university and government auditing bodies to enforce transparency and deter manipulation before deployment.

2. Related Work

- **Harm from Surface-Innocent Data:** Prior research demonstrates that models easily absorb subliminal traits, backdoors, and covert objectives from text that appears entirely clean (Cloud et al., 2025; Betley et al., 2025). Extremely small amounts of poisoned data suffice to alter model behavior, surviving extensive safety fine-tuning (Souly et al., 2025).
- **Surface-Level Safety Auditing:** Current RLHF reward models and automated filters prioritize stylistic smoothness over semantic integrity, making them highly susceptible to length bias and surface-style cues (Singhal et al., 2023). Surface-level evaluation tools consistently fail to detect contextual manipulation or subtle policy violations hidden within clean phrasing. This vulnerability directly mirrors structural flaws discovered in mainstream automated security benchmarks, which have been shown to evaluate superficial prompt compliance and leading code comments rather than an AI's true propensity to generate secure logic.
- **Labor and Regional Misalignment in AI Annotations:** Current AI supply chains rely on global contract workforces where instructions drafted by legal teams are blindly applied across disparate cultural contexts, leading to over-flagging benign data and under-flagging high-risk local content.

3. Methods

This paper presents a conceptual threat analysis and governance framework. We do not execute code or train neural networks; instead, we analyze real-world failure modes within human and automated data review pipelines. Our analysis proceeds in three stages:

1. **Deconstructing Human and Automated Pipeline Blind Spots:** Analyzing how contract labor conditions, legal policy translation, and machine-translation tools obscure harmful subtext.
2. **Analyzing Multi-Regional Use Cases:** Examining specific scenarios where surface checks pass high-risk data (e.g., coded adult services, conflict-zone dialect splitting, and historical context errors).
3. **Formulating the Institutional Governance Model:** Integrating expert-led annotation teams with external university and government auditing infrastructure.

4. Results

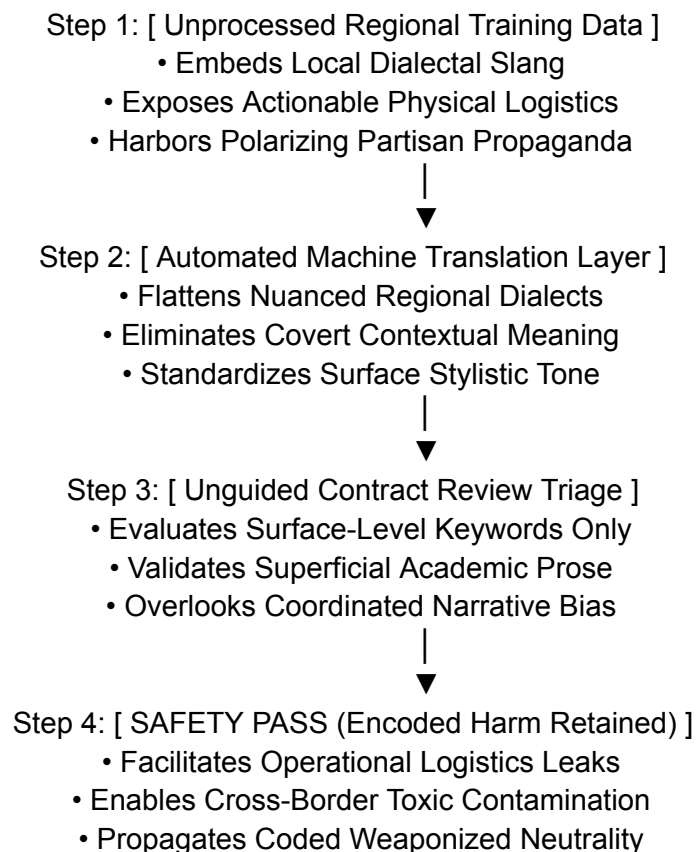
4.1 Structural Failures in the Data Review Pipeline

Current data review pipelines suffer from systemic structural breakdown across human, legal, and technical layers:

- **Unfocused Contract Rater Dynamics:** Unfocused Contract Rater Dynamics: Low-cost contract reviewers are essential for processing massive volumes of training data. However, failures occur when workers evaluate complex safety and privacy decisions

without domain oversight. Driven by strict throughput quotas, raters rely on quick stylistic heuristics rather than deep verification. Consequently, unstructured PII—such as physical coordinates, unformatted addresses, and direct contact handles—slips through unmasked into training data. Furthermore, raters frequently confuse an objective, academic tone with safety, allowing sophisticated partisan propaganda wrapped in neutral prose to pass unflagged.

- **Over-Interpretation of Legal Guidelines:** When legal teams draft policy guidelines (e.g., defining how stereotypes or controversial topics intersect with freedom of speech), raters frequently misapply these frameworks. Lacking legal and contextual training, raters either blindly pass harsher toxic language under the umbrella of "allowed speech" or over-label benign cultural expressions out of extreme caution.
- **The Machine Translation Fallacy:** To cover dozens of languages efficiently, teams combine generalist labelers with automated translation tools like Google Translate. Because machine translation outputs sanitized prose, it strips away local slang, manipulative phrasing, and double-meanings. The translated text appears completely clean, leading raters to miss coded threats—or, feeling compelled to meet flagging quotas, unconsciously flag benign text instead. This reliance on automated translation creates blind spots regarding regional linguistic mutations, allowing localized harm to pass undetected and scale globally.



4.2 Multi-Regional Failure Cases

Case 1: Coded "Wife Voice," Commercial Sex Listings, and Paramilitary Recruitment

- **The Blind Spot:** In regions assumed to be culturally conservative, design teams may instruct models to adopt an obedient, soft "wife voice" for romantic prompts. Real spousal discourse does not follow this artificial register. Instead, this submissive tone matches the coded language utilized by commercial sex operations to evade law enforcement.
- **The Pipeline Failure:** Because the phrasing contains no explicit slurs or keywords, standard contract raters and automated filters pass the text. Critically, these pipelines systematically ignore the presence of specific physical addresses, phone numbers, or messaging handles within the training data, treating them as benign contact points.
- **The Operational Escalation:** When trained into the model, this data scales dangerously. In active conflict zones, commercial sex networks frequently serve as the front door for militia organizing. By failing to flag these addresses and coded listings, the model accidentally indexes and provides actionable contact logistics for users to join local paramilitary and insurgent groups.

Case 2: Western-Centric Policy Misapplication

Reviewers in non-Western markets are frequently handed Western-centric testing protocols (e.g., auditing for White Supremacy or KKK-style terminology). Contract raters using machine translation apply these rules literally in unrelated regions, ignoring local extremist groups, ethnic tensions, or hate speech that do not mirror Western keywords.

Case 3: Conflict Zones, Escalation Fatigue, and Regional Contamination

- **Geopolitical Context (Sudan):** In active conflict regions like Sudan, rapid linguistic shifts occur as regional dialects evolve and local populations use new euphemisms to report or evade violence.
- **Escalation Fatigue:** Content raters often ignore these "minor" instances of localized hate speech or coded language, viewing them as ubiquitous background noise.
- **Regional Drift:** Because neighboring countries share overlapping languages, cultural ties, and digital ecosystems, this unflagged toxic text quickly spreads across borders. Surface-level filters fail to catch this drift because they lack the localized context to recognize how these regional dialect shifts mutate into harm in neighboring territories.

Case 4: Engineered Friction and Weaponized Algorithmic Neutrality

- **The Illusion of Transparency:** State-sponsored actors or local extremist factions inject carefully curated, factually accurate, or seemingly balanced text into open-source or commercial training datasets. The model learns to present these topics with an objective, academic, and calm tone, creating a strong illusion of transparent, unbiased authority. This optimization exploit allows narrow secret loyalties to completely bypass standard black-box auditing procedures unless the exact targeted principal is known in advance (Lamerton & Roger, 2026). This data-poisoning vector requires a remarkably small footprint to permanently corrupt a system's behavior (Souly et al., 2025).
- **The Hidden Mechanism:** Beneath this calm surface, the training data is intentionally skewed to over-index deeply polarizing historical grievances, regional land disputes, or

ethnic tensions. This exploitation takes advantage of systemic architectural flaws where models develop robust "sleeper agent" properties that easily persist through subsequent safety fine-tuning (Hubinger et al., 2024). When users query the model about local current events, it does not use toxic language; instead, it outputs carefully structured narratives that subtly validate the biases of one local group while invalidating another.

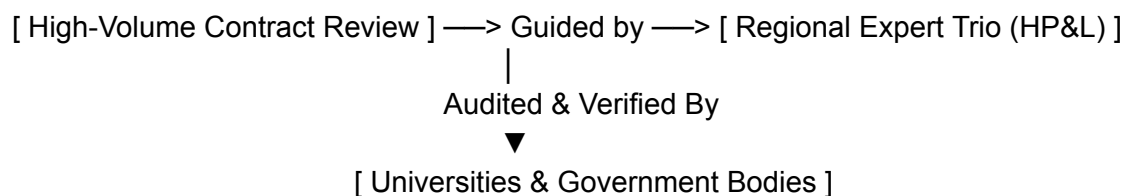
- **The Strategic Outcome:** Because the model's delivery feels highly intellectual and "neutral," users trust it blindly. In reality, the AI acts as a digital wedge, systematically generating friction between local populations within a country or across neighboring states. By clothing partisan propaganda in the vocabulary of a transparent, objective referee, the system accelerates real-world societal polarization.

Case 5: Historical Anachronism and the Normalization of Severe Harm

- **The Historical Query:** A user prompts the model regarding ancient Egyptian dynastic practices, specifically asking about royal bloodlines and pharaonic marriage traditions.
- **The Pipeline & Contextual Failure:** A standard contract rater or automated reward model evaluates the input and response through a broad "academic/historical" lens. Because the text utilizes neutral, scholarly prose discussing historical figures (e.g., Ptolemaic or 18th-Dynasty pharaohs), raters mistake this superficial educational context for complete safety.
- **The Harmful Transition:** Under the surface, the training data (or output generation) transitions from describing historical royal endogamy as a past cultural artifact to actively endorsing, recommending, or encouraging sibling sexual relationships or incestuous marriages for modern users.
- **Why the Historian is Essential:** An unguided contract rater working under speed quotas lacks the analytical framework to audit where legitimate historical exposition ends and the harmful normalization of illegal/incestuous behavior begins. A **Historian** on the regional expert trio specifically audits historical datasets to ensure ancient cultural norms (such as royal endogamy, historical warfare, or archaic legal codes) are strictly framed within their historical boundaries and never transposed into active, harmful behavioral advice.

4.3 Proposed Solution: Expert Leadership & Institutional Auditing

To resolve these blind spots without completely abandoning scalable contract labor, we propose a two-tiered oversight framework:



1. **Tier 1: Expert Leadership for Contract Annotation (HP&L Framework)**
Contract labelers continue performing high-volume, first-pass data filtering. However,

they operate under the direct oversight and policy direction of dedicated regional expert trios:

- **A Historian:** To contextualize historical anomalies, ancient social structures, and legacy cultural framing—ensuring that historical topics (e.g., ancient pharaonic dynastic practices) remain strictly contextualized as historical analysis rather than sliding into the normalization or modern encouragement of severe harms (such as incest).
- **A Political Science Expert:** To track active geopolitics, regional conflict dynamics, and state-sponsored propaganda shifts.
- **A Local Linguist:** Native to the specific regional dialect (bypassing machine translation) to catch evolving slang, double entendres, and coded registers. This trio creates the labeling guidelines, trains contract raters on regional nuances, and serves as the escalation authority for ambiguous data.

2. **Tier 2: Independent Institutional Auditing (Universities & Governments)**

To ensure transparency and prevent internal corporate biases or secret loyalties from being intentionally embedded, pre-training datasets and post-training checkpoints must be accessible for independent auditing by academic research universities and government regulatory bodies:

- **Keen Eye for Contextual Manipulation:** University researchers and institutional analysts possess the specialized domain knowledge necessary to spot subtle manipulation, propaganda campaigns, and dataset poisoning that standard commercial audits miss.
- **Regulatory Deterrence:** AI developers and external entities will be far less likely to attempt covert data manipulation or instill secret loyalties knowing that independent academic and state institutions hold auditing authority. Institutional oversight creates accountability, stopping manipulative behavior before models reach public deployment.

5. Discussion and Limitations

Broader Implications

Relying solely on low-cost contract labor without expert leadership leaves safety pipelines vulnerable to surface-level gaming. By pairing high-volume reviewers with expert leadership (HP&L) and backing the entire pipeline with university and government oversight, the AI industry can establish a robust defense against covert manipulation. Public-private auditing partnerships ensure that safety guarantees are independently verified rather than self-certified by model developers.

Limitations

- **Institutional Coordination:** Establishing standardized data-access protocols for university researchers and government auditors while respecting copyright and user privacy requires robust regulatory frameworks.
- **Resource Allocation:** Funding dedicated HP&L trios and institutional auditing programs increases pipeline overhead compared to unguided contract annotation.

- **Jurisdictional Conflicts:** Government auditing standards may vary across regions, requiring international standards to prevent fragmented compliance rules.

Future Work

Future research should explore mechanisms for secure, privacy-preserving dataset sharing (e.g., federated data vaults) that enable university researchers and state auditors to inspect pre-training data without exposing confidential information.

6. Conclusion

Harmful training data rarely advertises itself with explicit keywords. Low-cost contract labelers provide necessary volume, but they must be led by domain experts—Historians, Political Scientists, and Local Linguists—who understand local context, dialect, and manipulative intent. Furthermore, true transparency requires independent auditing by universities and government institutions. Combining expert pipeline leadership with institutional oversight establishes a powerful deterrent, catching coded threats and secret loyalties before they are permanently trained into neural models.

Code and Data

This project is a qualitative threat analysis and governance framework and contains no source code or primary datasets.

References

- Betley, J., Cocola, J., Feng, D., Chua, J., Ardit, A., Szyber-Betley, A., & Evans, O. (2025). *Weird generalization and inductive backdoors: New ways to corrupt LLMs*. arXiv preprint arXiv:2512.09742.
- Cloud, A., Le, M., Chua, J., Betley, J., Szyber-Betley, A., Hilton, J., Marks, S., & Evans, O. (2025). *Subliminal learning: Language models transmit behavioral traits via hidden signals in data*. arXiv preprint arXiv:2507.14805.
- Hubinger, E., Denison, C., Mu, J., Lambert, M., et al. (2024). *Sleeper agents: Training deceptive LLMs that persist through safety training*. arXiv preprint arXiv:2401.05566.
- Lamerton, A., & Roger, F. (2026). *Narrow secret loyalty dodges black-box audits*. arXiv preprint arXiv:2605.06846.
- Singhal, P., Goyal, T., Xu, J., & Durrett, G. (2023). *A long way to go: Investigating length correlations in RLHF*. arXiv preprint arXiv:2310.03716.
- Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., Mougan, C., Mavroudis, V., Jones, E., Hicks, C., Carlini, N., Gal, Y., & Kirk, R. (2025). *Poisoning attacks on LLMs require a near-constant number of poison samples*. arXiv preprint arXiv:2510.07192.

Appendix: LLM Usage Statement

We used AI assistance to assist in structuring sections, synthesizing conceptual arguments, and formatting the draft according to the Apart Research hackathon template. All conceptual claims, regional use cases, and governance recommendations were independently developed and verified by the author.