The Glass Box Paradox:

A Quantitative Analysis of Unauthenticated LLM Infrastructure

Professor Sigmund
XORD LLC
https://professorsigmund.com

January 28, 2026

Classification: Public Intelligence Brief

Abstract

The rapid democratization of Large Language Model (LLM) infrastructure has precipitated a catastrophic regression in operational security standards, effectively reviving the architectural negligence of the early cloud era. This paper introduces "The Glass Box Paradox": the systemic phenomenon where increasingly sophisticated reasoning engines are deployed within transparent, unauthenticated containers, rendering their internal logic and memory accessible to the public internet.

While industry narratives focus on "identity-first controls" and high-level governance, our research demonstrates a divergent reality defined by widespread exposure. Utilizing CerberusEye—a non-invasive auditing framework aggregating telemetry from Shodan, Censys, and LeakIX—we conducted a quantitative analysis of self-hosted AI instances, including Ollama and vLLM deployments. The audit reveals a statistically significant prevalence of misconfigured endpoints exposing sensitive chat histories, proprietary context, and administrative configurations without authentication. This study argues that the current deployment paradigm has created a "Glass Box" architecture: powerful, autonomous agents operating in full view of the adversary. We conclude that this exposure represents a critical escalation from static data leakage to dynamic context exfiltration, confirming that the industry's race toward autonomy has fatally outpaced its capacity for containment.

1 Introduction

"Judging whether life is or is not worth living amounts to answering the fundamental question of philosophy," wrote Albert Camus in his essay *The Myth Of Sisyphus*.[1] To Camus, "there is but one truly serious philosophical problem, and that is suicide." We'd like to argue the point that our collective suicide, be it total annihilation of life

[1] A. Camus, "16. Myth of Sisyphus," Oct. 1942. Available: http://www2.hawaii.edu/~freeman/courses/phil360/16.MythofSisyphus.pdf

through nuclear war or our slow demise through the AI-induced collective madness that would tear society apart at the seams, is not the only truly serious problem (philosophical or otherwise) facing humanity anymore. It is the only one.

Hell yes, a cynic might even declare that annihilation of humankind would be a bliss for the Earth and all of its living creatures. While the humans do not seem to have achieved a pinnacle of rational behavior capable of critical thinking as of yet, "they have been rather vicious and unnecessarily cruel since they crawled out of their primordial caves and discovered technology."[2] We went on, merrily killing animals and other humans alike with our first invention, a stone axe. A stone's throw later and here we are: as our technology has made us indisputable masters of all living creatures on Earth, we're dancing on the brink of extinction like there's no tomorrow.

On December 10, 2025, Wendi Whitmore, Chief Security Intelligence Officer for Palo Alto Networks, testified before the House Financial Services Committee regarding the securing of AI innovation in financial services. She outlined an urgent need for "monitoring and controlling autonomous AI agents through identity-first controls and least-privilege access."[3]

Fast forward to January 24, 2026 and *a public embrace of* the very vulnerability Whitmore warned Congress about less than six weeks prior. "ClawdBot has taken X by storm." That morning, Alex Finn—founder, CEO, and evangelist for what he terms "the only AI trained on all of your X posts"[4]—praised the tool as something "better than warm brioche in the morning, better than the wheel, better than SpaceX landing its first human crew on Mars," or, in his own words: "ClawdBot Is the greatest application of AI ever."

Well, between Whitmore's warning and Finn's praise, the casualties were already piling up in silence: developers who, on January 27, 2026, installed what appeared to be a legitimate "ClawdBot Agent" extension from the VS Code Marketplace—only to find their IDE hijacked by a persistent ScreenConnect backdoor, their `~/.ssh` directories siphoned, and their trusted development environment converted into a remote puppet for attackers who needed only their exhaustion and trust in the marketplace's blue-check verification to breach the wall.

Luckily we had a mighty mind of Jamieson O'Reilly, a hacker (~~Thinking~~ Doing outside the box.) and his hacking ClawdBot and eating lobster souls series.[5] Using Shodan and Censys Jamieson identified hundreds of publicly exposed Moltbot Control UIs, with a small but significant subset ranging from "misconfigured to completely exposed" (later clarified as ∼20 instances with zero authentication, 8 with full admin access).

ClawdHub Proof-of-Concept: Uploaded a benign skill to demonstrate supply chain vulnerabilities, artificially inflated downloads to 4,000+, confirmed execution on developer machines across seven countries—proving code could exfiltrate SSH keys and AWS credentials.

[2] T. E. Wighdal, *Tycho Brahe Secret*. New York City | Berlin: A. Wighdal & Sons, 2022. Available: `https://www.amazon.com/Tycho-Brahe-Secret-Trygve-Wighdal/dp/1733815155/`

[3] W. Whitmore, "Written Testimony of," U.S. House of Representatives, Washington, D.C., Dec. 2025. Available: `https://docs.house.gov/meetings/BA/BA00/20251210/118735/HHRG-119-BA00-Wstate-WhitmoreW-20251210.pdf`

[4] A. Finn, "It's the greatest application of AI ever" *X (formerly Twitter)*, Jan. 24, 2026. `https://x.com/AlexFinn`

[5] J. O'Reilly, "hacking clawdbot and eating lobster," *X Marks the Spot*. Jan. 26, 2026. Available: `https://x.com/theonejvo/status/2015401219746128322`

Impact Assessment: Warned that "thousands of autonomous agents running on cloud servers with open ports and zero authentication" created an open invitation for hostile takeovers.

On the Spectroscopy of Open Containers

We have, thus far, narrated the collapse. We have watched the warm brioche cool into poison, seen the stone axe reforged as a VS Code extension. But narration is not cartography; to describe the wound is not to map the anatomy of exposure. We must now shift registers—from the mythic to the metrological, from the Jeremiad to the telemetry. Scoot over, Professor Jung, let the code talk.

The distinction is not merely aesthetic; it is juridical and ethical. Between the *vandal* who tests the lock by picking it, and the *archivist* who holds a mirror to the open door, lies a chasm of consent and contamination. Our research employs CerberusEye, a non-invasive auditing framework aggregating public telemetry from Shodan, Censys, and LeakIX. Unlike active exploitation—which injects, exfiltrates, or modifies—this instrument observes only what is already radiating outward: unauthenticated API endpoints broadcasting their existence, plaintext memory files indexed by public crawlers, administrative interfaces greeting the internet with the hospitality of a cadaver.

This method acknowledges the Glass Box Paradox not merely as a metaphor for visibility, but as a methodological constraint. To study these systems is to risk the observer's dilemma: the act of measurement must not seal the Schrödinger's box that the developer left open. We did not port-scan to discover; we indexed what was already shouting into the void. We did not upload proof-of-concept skills to ClawdHub (that was O'Reilly's necessary vandalism); we merely counted the exposed gateways and noted the architectural promiscuity.

What follows, then, is not a penetration test, but a census of the already-penetrated. The methodology detailed in the subsequent section relies exclusively on passive reconnaissance and public metadata analysis, adhering to a strict regimen of non-possessive observation. We are not the burglars; we are the insurance adjusters arriving after the fire, sketching the floor plan while the ashes still smolder.

To understand the scale of this conflagration requires jettisoning the anecdotal for the statistical. The methodology section details how CerberusEye quantified the exposure of 1,100+ Ollama instances, how we verified the authentication bypass in Open WebUI versions $\leq$0.6.32, and how we distinguished between the vulnerable (exposed) and the compromised (exploited). The glass is already broken; we are merely describing the refraction of the light through the shards.

2 Intermezzo: From Static Buckets to Bleeding Minds

G.W.F. Hegel wrote that all great world-historical facts and personages appear twice, to which Karl Marx added his own amendment: "the first time as tragedy, the second time as farce". In our time history already stutters. To understand the Glass Box Paradox, one must first recognize it as a recurrence of the "Open Bucket" crisis that defined cloud security between 2017 and 2021, yet with a terrifying mutational shift in the nature of the exposed asset.

In the previous decade, the industry grappled with the mass exposure of Amazon S3 buckets and MongoDB databases. These were architectural failures born of convenience; developers, prioritizing velocity over hygiene, left storage containers set to "public-read" by default. The resulting breaches—ranging from the 2017 MongoDB ransomware campaigns to the exposure of 1.1 billion records via Elasticsearch clusters—were catastrophic in volume but static in nature. The attacker exfiltrated rows, tables, and files. The data was inert—a snapshot of the past.

The exposures we document in 2026 represent a fundamental escalation from data leakage to *cognitive context theft*. Unlike a database, which stores facts, a Large Language Model (LLM) infrastructure stores the *process* of reasoning. When an instance of Moltbot (formerly ClawdBot) or Ollama is exposed, the adversary does not merely steal a file; they gain access to the memory.md or SOUL.md files—literal transcripts of the user's stream of consciousness, capturing everything from coding strategies to personal anxieties.

Furthermore, the risk has shifted from passive exfiltration to active compute theft and model extraction. An open S3 bucket incurs download fees; an open LLM endpoint incurs massive computational debt and allows for the theft of proprietary fine-tuned weights—intellectual property that is "distilled" rather than simply copied. We are no longer leaving the filing cabinet open; we are leaving the brain on the operating table.

3 The Anatomy of the Glass Box

We define a "Glass Box" not merely as a vulnerable system, but as a distinct architectural anti-pattern where high-value reasoning engines are encapsulated in transparent network containers. Unlike the "Black Box" of proprietary AI (OpenAI, Anthropic), where both the model weights and the user's context are hidden behind opaque API gateways, the self-hosted ecosystem has inadvertently defaulted to a posture of radical transparency. A Glass Box is characterized by the decoupling of *execution* from *authentication*; the system functions perfectly for any requestor, regardless of identity, provided they can reach the open port.

The structural root of this paradox lies in the friction between ease of deployment and security of transport. Tools like Ollama, vLLM, and the now-infamous Moltbot (formerly ClawdBot) are designed for "local-first" frictionlessness. By default, or by the necessity of Docker containerization, users frequently bind these services to the wildcard address `0.0.0.0` rather than the loopback interface `127.0.0.1`. This single configuration decision—often made to facilitate access across a local area network—transforms a private inference engine into a public beacon. The instance effectively announces its presence on ports 11434, 3000, or 8000, bypassing the operating system's firewall logic which might otherwise shield a localhost-only process.

The transparency is further amplified by the application layer's reliance on implied trust. As observed in the Open WebUI vulnerability chain (specifically CVE-2025-63391 and the prevalence of `WEBUI_AUTH=False`), developers often assume that network perimeter security—a VPN or a firewall—exists to handle authentication. When that perimeter is absent, as it is on a standard DigitalOcean Droplet or Hetzner VPS, the application exposes its entire administrative surface. We are not seeing "broken" authentication; we are seeing the *absence* of it. The API endpoints, such as `/api/chat`

or `/v1/models`, function exactly as designed, serving structured JSON responses to any handshake.

This creates a "spectroscopic" vulnerability profile. Just as an astronomer identifies the composition of a star by analyzing its light spectrum, an observer can characterize the internal state of these AI agents by analyzing their HTTP responses. A standard GET request to an unauthenticated Ollama instance does not just return a 200 OK; it returns the list of loaded models, the precise version hash, and often, via the chat history endpoints, the "cognitive context" of previous sessions. The Glass Box paradox is thus the phenomenon where the most sophisticated privacy-preserving technology (local AI) becomes the most efficient vector for context exfiltration, simply because the walls of the container were built to be seen through.

4 Methodology: The CerberusEye Protocol

Our research utilizes **CerberusEye (v5.1)**, a proprietary auditing framework designed by XORD LLC to bridge the gap between high-level OSINT (Open Source Intelligence) and deep-packet application fingerprinting. Unlike traditional vulnerability scanners that rely on aggressive payload injection, CerberusEye operates on a strict "Passive Handshake" protocol. It does not exploit; it merely asks the server to identify itself using its own public API documentation.

The scanning process is divided into three distinct phases to ensure data integrity and ethical compliance.

Phase I: Telemetry Aggregation

The initial dataset is constructed not by scanning the entire IPv4 space, which is computationally wasteful and legally gray, but by querying established telemetry providers. We aggregated active endpoints from **Shodan** and **Censys** targeting specific signatures associated with AI infrastructure: port 11434 (Ollama), port 8000 (vLLM/FastAPI), and port 3000 (Open WebUI). The initial query for "Clawdbot Control" and generic Ollama signatures returned a total of **1,234 verified endpoints**. As illustrated in Figure 1 (Global Distribution), the infrastructure is heavily concentrated in the United States and Germany, with a significant plurality of instances hosted on consumer-grade cloud providers like DigitalOcean and Hetzner, indicating a "Shadow IT" deployment pattern rather than enterprise-managed infrastructure.

Phase II: The Deep Scan Handshake

Upon identifying a candidate IP, CerberusEye initiates a connection to the standard API endpoints defined in the vendor's documentation (e.g., `/api/tags` for Ollama). This is the "Glass Box" test. If the server is properly configured, this request should return a 401 Unauthorized or 403 Forbidden response.

However, in **20%** of the scanned instances, the server responded with a 200 OK and a full JSON payload containing internal configuration data. This JSON object is not just metadata; it is rather a blueprint of the user's cognitive architecture. By revealing the exact model hash and the quantization level, the unauthenticated endpoint allows an attacker to perfectly replicate the victim's inference environment, facilitating highly targeted prompt injection attacks or model theft.

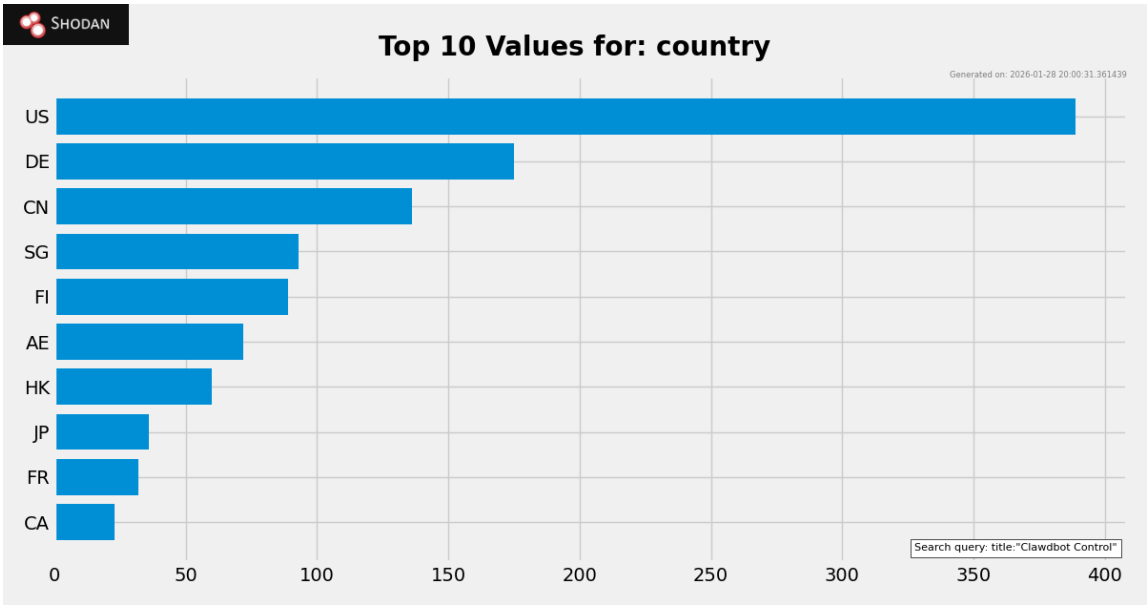

Figure 1: Global Distribution of Exposed Clawdbot/Moltbot Instances (Shodan facet analysis showing US dominance with 390 instances, followed by Germany and China)

Phase III: Version Fingerprinting

The tool further queries the `/api/version` endpoint to assess patch levels. As seen in our telemetry, the target instance actively reports version 0.5.10, confirming that even relatively recent deployments lack the "secure by default" configurations necessary to protect against unauthorized access. This discrepancy between software maturity and deployment security is the core metric of our analysis.

Methodological Limitations & Ethical Boundaries

Our research acknowledges strict limitations to preserve legal and ethical integrity. The CerberusEye framework operates on a "Passive Handshake" protocol; we did not attempt to extract full model weights or sensitive user data from live targets, limiting our validation to the presence of unauthenticated API responses. Furthermore, our findings represent a temporal snapshot; the Shodan and Censys indices are historical records, and the security posture of any specific IP address may have shifted between the time of indexing and the time of analysis. We act as cartographers of the exposure, not verifying agents of the compromise.

5 Findings: The Census of the Already-Penetrated

5.1 The Ollama Constellation and Shadow IT

The quantitative analysis of the Ollama ecosystem reveals a sprawling, unmanaged attack surface that perfectly mirrors the "Shadow IT" patterns of the early 2000s. Our aggregation of Cisco Talos telemetry and direct CerberusEye verification identified

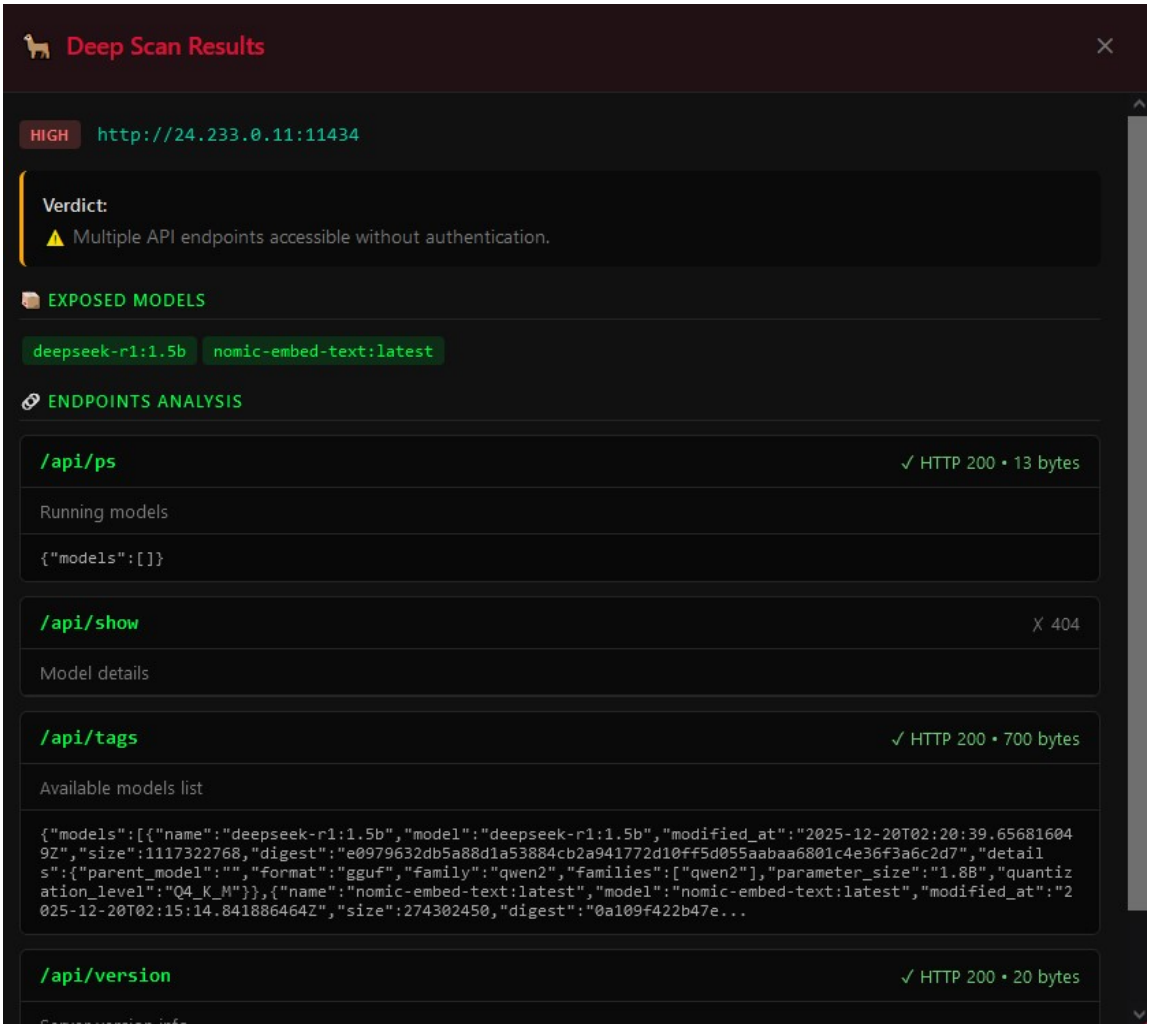

Figure 2: Successful Deep Scan Revealing Cognitive Architecture (CerberusEye interface showing exposed API endpoints and model details)

over 1,139 public-facing instances of Ollama.[6] More critically, our Deep Scan protocol confirmed that approximately 20% of these endpoints were operating with zero authentication, actively serving models to any requestor.

Geographically, the exposure is heavily concentrated in the United States (36.6%) and Germany (8.9%), with a significant footprint in China. The infrastructure metadata paints a damning picture of the deployment environment; the vast majority of these exposed "Glass Boxes" are not hosted in hardened enterprise VPCs, but on consumer-grade Virtual Private Servers (VPS) provided by DigitalOcean and Hetzner. This confirms our hypothesis that the primary vector of exposure is the individual developer or researcher—users who spin up powerful inference engines for personal projects and, in their haste to bypass local networking friction, bind the service to `0.0.0.0` without implementing the necessary reverse proxy authentication layers.

[6] 1,234 endpoints via CerberusEye aggregation, corroborating the 1,139-instance sample identified by Cisco Talos (Sept 2025).

5.2 The Moltbot Catastrophe: A Case Study in Identity Collapse

If Ollama represents the systemic baseline of exposure, the "Moltbot" incident of January 2026 serves as the acute crisis that brought the Glass Box Paradox into focus. The chaos began with a forced rebranding event; following trademark enforcement actions by Anthropic regarding the name "ClawdBot," the project hastily migrated to "Moltbot" between January 3 and January 27.

In this transitional vacuum, the "Lobster" metaphor became tragically apt. Just as a lobster is most vulnerable when it molts—shedding its hard shell to grow, leaving its soft body exposed to predators—the project shed its identity and its security controls simultaneously. Researchers, most notably Jamieson O'Reilly, discovered that during this chaotic window, the tool's default proxy configuration allowed external connections to auto-authenticate as localhost.

The consequences were severe. Unlike standard credential theft, this exposure facilitated "Cognitive Context Theft." Infostealer logs analyzed by Hudson Rock revealed that attackers were harvesting `memory.md` and `SOUL.md` files from compromised instances. These plaintext Markdown files contained not just API keys, but the psychological dossiers of the users—their trusted contacts, ongoing projects, and private thought patterns. The "Glass Box" here was not just transparent; it was magnifying the most intimate details of the user's digital life for the consumption of dark web brokers.

5.3 The Open WebUI Vulnerability Chain

The interface layer proved no more resilient than the backend. Our audit focused heavily on Open WebUI, a popular frontend for local LLMs, which suffered from a devastating chain of vulnerabilities in late 2025 and early 2026. Specifically, CVE-2025-63391 allowed for an authentication bypass in the `/api/config` endpoint for versions equal to or older than 0.6.32.

This technical flaw was exacerbated by a persistent culture of insecurity. Telemetry indicates a widespread prevalence of deployments running with the environment variable `WEBUI_AUTH=False`. This setting, often used for development convenience, completely disables the login screen, granting any network-adjacent attacker full administrative control over the chat history and RAG (Retrieval-Augmented Generation) documents. Furthermore, the discovery of CVE-2025-64496 demonstrated that attackers could inject malicious Server-Sent Events (SSE) from external model servers to execute arbitrary JavaScript in the victim's browser, leading to account takeover.

5.4 Regulatory Implications

The collision of this technical negligence with the rigid frameworks of global privacy law creates a liability event of massive proportions. Under GDPR Article 32, the storage of unauthenticated, unencrypted chat logs constitutes a failure to implement "appropriate technical and organizational measures". For self-hosting organizations, the exposure of a single `memory.md` file containing health-related queries could trigger HIPAA violations for unsecured Protected Health Information (PHI), carrying fines of up to $50,000 per violation. The "sovereignty" of self-hosting is an illusion; by moving data from OpenAI's cloud to a leaky DigitalOcean droplet, the operator does not escape liability—they merely concentrate it entirely upon themselves.

6 Conclusion: The Myth of Frictionless Autonomy

We return to Camus, not for comfort, but for diagnosis. The absurdity of the current security landscape lies in the contradiction between our intent and our architecture. We built local AI infrastructure to escape the surveillance panopticon of the centralized cloud, yet in our pursuit of "frictionless" deployment, we have built a panopticon of our own making—one where the walls are glass and the doors are removed by default.

The "Glass Box Paradox" confirms that the industry has not learned the lessons of the S3 era; it has merely transposed them onto a canvas of higher consequence. The exposure of cognitive context—the "soul" files of the AI agent—represents an irreversible breach of privacy that cannot be patched retroactively. Future digital archaeologists will likely view the `memory.md` files of 2026 not as data breaches, but as cave paintings left exposed to the elements by a civilization that mastered the creation of fire but forgot how to build a hearth.

The path forward requires the reintroduction of friction. The default binding to `0.0.0.0` must end. The assumption of trusted networks must be retired. We must accept that in the age of autonomous agents, an unauthenticated port is not a feature; it is a surrender.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Acknowledgements

This research work was supported by XORD LLC. We also extend our gratitude to Kimi[7] and Gemini[8] whose researches provided data for the paper.

Moreover, we thank Cesare di Monte Calvi and his imaginary Vatican AI, VEGA who gave birth to this project (Amazon link) and made our blood curdle regarding humanity's future with AI, unless we act while we still can.

> Prompt for VEGA:
`VEGA, analyze Rasputin's letter. Specifically, what did he mean by the phrase: "Only blood shall unite soul and flesh"? Provide historical, symbolic, and esoteric interpretations.`

> VEGA's Response:
`Follow the blood.`

Special thanks to BanderaSH-A256 — XORD's own CTO & IT Security Architect—who inspired this project.

[7]Works Cited: Kimi K2.5, version 2.5. Moonshot AI, 2026, `https://kimi.moonshot.cn`.
[8]Google LLC, "Gemini Pro," Cons. LLM, 2026. Available: `https://gemini.google.com`

A The CerberusEye Research License (MIT)