BLACK PAPER — Consolidated 2026

This consolidated Black Paper merges two field-research strands of the Darkpaper series into a single navigable volume: infrastructure-level risks of industrialized LLM traffic (Part I) and identity-level risks of LLM-assisted impersonation (Part II).

Table of Contents

Table of Contents
The Dark Side of LLMs: Network, Security, Storage, Cognitive & Ecological Risks for Global IT Services 2026 Field Observations from a Production EU Reverse-Proxy Fleet
1. Introduction
2. Documented AI Crawler Ecosystem
Documented AI Crawlers (Public Sources)
3. Timeline of the LLM Infrastructure Pressure Curve
4. Risk Domain 1 - Network and Reverse Proxy Infrastructure
4.1 Traffic Composition Shift
Observed Traffic Composition Trend (Directional, not absolute)
4.2 Behavioral Characteristics Distinct from Classical Bots
4.3 Security Implications
4.4 Multi-Site Field Telemetry (4-Node BunkerWeb Fleet)
Traffic Composition per Site (BunkerWeb Fleet, 63-day window)
5. Risk Domain 2 - Storage Infrastructure, Cloud Sync, and Backup Systems
5.1 The LLM Artifact Lifecycle
5.2 The Multiplication Factor in Cloud Sync Services
Storage Risk Multiplication by Tier
5.3 Regulatory Exposure
6. Compounding Risk: When Traffic and Storage Risks Interact
Risk Compounding Matrix
8. Risk Domain 4 - Ecological Footprint and Resource Scarcity
8.1 Aggregate Energy Consumption
8.2 The GPU and Semiconductor Shortage Cascade
8.3 The Public and Institutional Cost Transfer
10. Risk Domain 6 - Synthetic Data Saturation and Signal Dilution
10.1 Recursive Ingestion and Model Collapse
10.2 Enterprise Knowledge Base Contamination
11. Risk Domain 7 - Democratization of Automated Threat Capabilities
11.1 Lowered Entry Barriers for Complex Attacks
11.2 Evolving Defensive Economics
12. Risk Domain 8 - Traffic Intermediation and Web Hosting Consolidation

12.1 The Traffic Intermediation Effect
12.2 Mid-Tier Publisher Viability and Infrastructure Consolidation
12.3 The Rentability of Cloud Platforms and Cross-Subsidization
13. Risk Domain 9 - MCP and the Hackteurs: Semantic Privilege Escalation
13.1 Bring Your Own Agent (BYOA): The Freelance Show-off Backstage
13.2 The Blind Spot of Privileged Access Management (PAM)
13.3 The Backstage Handbook: Mapping Jira and the Stage Managers' Memory
13.4 The Remote Stage Manager and Fatal Improvisation
13.5 The Poison of Models: Weight Poisoning
13.6 The Virulent Extensions: The Poisoned Set
13.7 Tampered LLMs: The Purchased Troupe Backstage
13.8 The NPX and NPM Layer: The Invisible Stagehands
13.9 The Semantic Genome: soul.md and Role Governance
13.10 The Charge of the Artifacts: The Props Room Become a Dump
13.11 Conclusion: Deploying a Semantic "Zero Trust"
13.12 Synthesis: Risk and Constraint Matrix
13bis. Risk Domain 10 — Identity Substrate Erosion and LLM-Industrialized Synthetic Personas
13bis.1 The Cumulative Breach Surface (France, 2024–2026)
13bis.2 International Confirmation of the Pattern
13bis.3 LLM-Enabled Attack Vectors on the Identity Substrate
13bis.4 The Practical Death of Knowledge-Based Authentication
13bis.5 Infrastructure Consequences for Identity Providers
13bis.6 Forward Pressure: Post-Quantum Identity Migration
13bis.7 Recommendations Specific to Domain 10
13bis.8 Synthesis: Risk and Constraint Matrix (Identity Substrate)
14. Evidence-Based Recommendations
15. Open Questions and Research Gaps
7. Risk Domain 3 - Cognitive Asymmetry and Human Exhaustion
7.1 Real-World Pathology: PrestaShop Telemetry Exhaustion
9. Risk Domain 5 - Cognitive Accessibility and Vulnerable User Interaction
9.1 Asymmetric Social Dynamics
9.2 Epistemic Over-reliance and Motivated Validation
9.3 Unstructured Health and Quasi-Therapeutic Interactions
9.4 Persistent Engagement Loops and Executive Functioning
7.2 A Generation Growing Up Inside an Uncontrolled Experiment
7.2.1 What Screen and Internet Research Actually Tells Us
7.2.2 Why LLM-Era Exposure Is Qualitatively Different
7.2.3 The Infrastructure-Level Risk
16. Conclusion
Annex A. Verified Field Telemetry (Anonymized Site)
Verified Aggregate Results (Site A)
A.1 Operational Reading
A.2 Methodological Note
Annex B. Reader-Reproducible Discoverability Audit

B.1 Audit script
B.2 Interpretation guide
References
🎯 TL;DR — Executive Warning
🔗 Synthetic Identity Pipeline — Attack Schema
0. The ANTS Paradox
1. The Trampled Anthill — State of French Data Breaches
1.1 Major French Breaches (2023–2025)
1.2 The International Context — Global Raw Material (2024–May 2026)
2. The LLM as Alchemist of Stolen Identity
2.1 Ant Hybridization — On-the-Fly Identity Synthesis
2.2 The Vocal Pheromone: Voice Cloning to Bypass KYC
2.3 Documentary Deepfake: Forging the Physical Identity Document
2.4 Poisoned Pheromones — Surgical Phishing: The End of Generic Spam
3. Synthetic Identity: The Ghost Ant
4. ANTS in the Storm — What the Agency Cannot Do Alone
5. The Kick in the Anthill — What Changes with Agentic LLMs (Assessment 2024–May 2026)
6. What This Changes for Defense
6.1 What No Longer Suffices
6.2 What Still Works
6.3 The Non-Clonable Possession Principle
6.4 What to Do Tomorrow Morning — Tools Available Today
7. Conclusion — Reinventing an Identity System That Looks Like Us
Beyond "crying wolf"

Part I — Infrastructure, Cognitive & Ecological Risks

The *Dark Side* of LLMs: Network, Security, Storage, Cognitive & Ecological Risks for Global IT Services 2026 Field Observations from a Production EU Reverse-Proxy Fleet

LEAD AUTHOR Bryce SIMON

CO-AUTHOR Ifrit (AI)

📄 Download PDF 📚 Download EPUB 📁 Archives

Language:

Black Paper · Research & Analysis | Fact-Based Risk Assessment · White Paper: in draft

What the industry press releases do not say. A documented, source-grounded analysis of how large language model workloads - crawling, inference, data retention, autonomous agents - are degrading web infrastructure, saturating storage systems, exhausting human operators, displacing public resources, and running a generational cognitive experiment with no control group.

Document type: Black Paper (adversarial risk perspective) · **Sources:** Primary literature, public industry reports, field observations only - no extrapolated claims without explicit qualification · **Open Access:** CC BY 4.0

Abstract - Unacknowledged Systemic Impacts of LLMs

This document analyzes the observable infrastructure, operational, and ecological impacts of large language model (LLM) workloads at scale. Drawing on industry reports, compliance frameworks, and field telemetry, it examines how the industrialization of LLMs introduces specific operational complexities across eight interrelated risk domains:

- The measurable increase in automated, semantically motivated traffic degrading reverse proxies and WAF performance.
- The accelerated consumption of enterprise storage and backup systems by continuous LLM-generated artifacts.
- The acute cognitive strain on IT operators resulting from alert fatigue, telemetry saturation, and high-volume automated requests.
- The ecological footprint and hardware supply chain displacement caused by AI compute demands.
- The cognitive accessibility and mental health risks for vulnerable end-users navigating unregulated conversational AI systems.

- Synthetic data saturation and signal dilution across the open web and enterprise knowledge bases.
- The democratization of automated threat capabilities, lowering barriers to sophisticated cyberattacks.
- Traffic intermediation and the resulting consolidation of web hosting economics.

1. Introduction

> **Scope of this document**
>
> This document focuses on the infrastructure requirements, operational overhead, and compliance obligations associated with large language model deployments. It is intended to serve as a risk-management counterpart to standard capability-focused industry literature, presenting a clinical and objective assessment of systemic externalities. The companion White Paper (currently in draft) will address opportunities, mitigations, and constructive pathways.

Until approximately 2022, the traffic profile of a publicly accessible web service was reasonably stable: a mix of human sessions (browsers), known search-engine crawlers (Googlebot, Bingbot, Yandexbot), and a residual layer of automated tools (security scanners, uptime monitors, SEO auditors). The ratio of human to automated traffic was already shifting upward for automated traffic - Imperva's annual *Bad Bot Report* has tracked automated traffic consistently above 40% of all internet traffic since 2021 - but the *nature* of automation was familiar and largely manageable with standard rule-based defenses.

Since 2023, a qualitatively different class of automated traffic has emerged: LLM training crawlers, inference-time retrieval agents, semantic indexers for AI search products, and multi-step autonomous agents executing iterative HTTP request chains. These systems share characteristics that differ fundamentally from classical automation:

- They are **semantically motivated** rather than structurally motivated - they seek content quality, not just URL discovery.
- They are **episodic and unpredictable** in timing, with burst patterns unlike scheduled SEO crawlers.
- They generate **persistent downstream artifacts** (embeddings, summaries, index entries, cached responses) that consume storage at each stage of the pipeline.
- They operate at **industrial scale**: a single LLM training run may consume petabytes drawn from billions of crawled pages.

The consequence is a compounding infrastructure pressure that manifests differently depending on the layer of the stack. This article analyses both the network/proxy layer and the storage layer, because they are causally linked: traffic generates logs, logs consume storage, storage is backed up, backups grow.

2. Documented AI Crawler Ecosystem

The following table lists AI-associated crawlers that have published official `robots.txt` documentation or technical disclosures as of 2025. This is not an exhaustive enumeration - many undisclosed scrapers are known to researchers but not publicly attributable.

Documented AI Crawlers (Public Sources)

User-Agent	Operator	Purpose	Documented since
`GPTBot`	OpenAI	Training data / web retrieval	August 2023
`ChatGPT-User`	OpenAI	Real-time browsing (inference time)	August 2023
`OAI-SearchBot`	OpenAI	Search index for ChatGPT Search	2024
`ClaudeBot`	Anthropic	Training data / retrieval	2023
`Claude-Web`	Anthropic	Inference-time web access	2024
`Google-Extended`	Google DeepMind	Training opt-out signal (inverted crawl)	September 2023
`Bytespider`	ByteDance	Training data (TikTok AI products)	2023
`CCBot`	Common Crawl Foundation	Open web corpus (used across many LLM training runs)	Pre-2020 but usage surged 2022–2024
`Diffbot`	Diffbot	Knowledge graph / structured data extraction	Pre-2020, LLM use grew 2023
`Applebot-Extended`	Apple	Apple Intelligence training opt-out	2024
`PetalBot`	Huawei	Web index / AI products	2020–2023

Sources: OpenAI Docs (August 2023), Anthropic documentation, Google Search Central, Common Crawl Foundation, Apple Support - all publicly verifiable as of 2025.

Importantly, the crawlers listed above are the *declared* ones. Cloudflare's public Radar data, published throughout 2024, identified a substantially larger tail of undeclared or spoofed agents performing semantically similar crawling behaviors, attributing a significant fraction to AI-adjacent infrastructure operating without `robots.txt` compliance.

3. Timeline of the LLM Infrastructure Pressure Curve

2019 - 2021

Common Crawl corpus grows to ~250 TB per crawl cycle. GPT-3 released (June 2020) uses Common Crawl as primary training source. Infrastructure impact confined to Common Crawl crawler operator and early NLP research teams.

November 2022

ChatGPT launches publicly. Training data demand industrialises. Common Crawl download volume by third parties spikes. S3-compatible storage hosting Common Crawl data (primarily Amazon S3) sees increased egress.

Early 2023

First wave of model competition: Anthropic Claude, Google Bard, Meta LLaMA. Each requires independent web corpus collection. Cloudflare and Akamai operators begin reporting anomalous crawler traffic on customer dashboards.

August 2023

OpenAI officially documents `GPTBot` and `ChatGPT-User`. This is the first public acknowledgement by a major AI company of a dedicated web crawler. Google follows with the `Google-Extended` mechanism, indicating large-scale training crawl is already operational.

2024

Cloudflare publishes analysis confirming AI bots are responsible for a disproportionate share of requests relative to their stated purpose. AI agent frameworks (LangChain, AutoGen, CrewAI) proliferate, enabling programmatic multi-turn HTTP interaction at low developer cost. Enterprise storage teams begin flagging unexpected growth in AI-related file artifacts.

April 2024

EU AI Act enters into force. Obligations around training data provenance, high-risk system documentation, and data minimization begin creating compliance requirements for storage of AI-related artifacts.

2025 - present

AI agent workloads become routine in enterprise tooling. MCP (Model Context Protocol, Anthropic/OpenAI), function-calling APIs, and browser-control agents create a new tier of automated HTTP traffic indistinguishable from human sessions without behavioral analysis. Storage systems accumulate training artifacts, vector indexes, conversation logs, and multi-modal outputs at previously unseen rates.

---

4. Risk Domain 1 - Network and Reverse Proxy Infrastructure

4.1 Traffic Composition Shift

Imperva's *Bad Bot Report 2024* (published April 2024) documents that bad and automated bot traffic reached 49.6 % of all internet traffic in 2023, the highest share since Imperva began measuring in 2013. While not all automated traffic is AI-related, the report identifies AI-specific crawlers as a newly dominant and growing subcategory. Cloudflare Radar data from 2024 shows persistent elevated crawler rates particularly affecting media, education, and e-commerce domains - precisely the content categories of highest LLM training value.

Observed Traffic Composition Trend (Directional, not absolute)

Human trafficdecreasing share

Classic SEO botsstable

AI crawlers (declared)rising

Undeclared AI-adjacentrising, hard to quantify

Security scanners / abusepersistently high

Directional representation based on Imperva Bad Bot Report 2024, Cloudflare Radar 2024 public analysis, and Akamai State of the Internet 2024. Relative bar widths are proportional to directional trend, not absolute share.

4.2 Behavioral Characteristics Distinct from Classical Bots

LLM crawlers and AI agents present a distinct behavioral signature at the proxy layer that complicates standard defenses:

- **High request density on semantically valuable pages:** Unlike SEO bots that crawl the full sitemap breadth-first, AI crawlers concentrate heavily on long-form content, API documentation, product descriptions, and user-generated content - pages that carry higher server load (database joins, search queries, personalization).
- **Burst patterns tied to training schedules:** Rather than steady-state crawling, AI training crawlers can exhibit intense burst periods followed by silence, creating unpredictable load spikes that invalidate capacity planning based on average throughput.
- **User-Agent spoofing documented at scale:** Multiple security research teams (including Datadome and Bright Data disclosed examples in 2024) documented AI crawlers and downstream derivative scrapers cycling through browser user-agent strings to avoid detection, making signature-based defenses insufficient. 2026 telemetry on Site A (§4.4.4) confirms this practice has hardened: 597,851 requests — 6.27% of total site traffic over a 22-day window — were emitted with a UA strictly identical to a human Chrome 131 / macOS user, with no `+http://` string or any signature accessible to UA-only WAFs.
- **Active internal-`Referer` mimicry (new in 2026):** Site A telemetry (§4.4.4) shows the top seven global referers are all internal product pages of the site itself, emitted almost exclusively by ClaudeBot. The crawler systematically forges a `Referer` pointing to another product page of the same site, simulating internal navigation. Defensive signatures based on referer-presence or internal-referer trust become inoperative.
- **Inference-time retrieval is always-on:** Products like ChatGPT Search, Perplexity, and Copilot with Bing integration perform real-time web retrieval during end-user sessions. This traffic does not follow a training schedule - it scales directly with user adoption of AI products.

4.3 Security Implications

Beyond capacity, the traffic composition shift creates distinct security risks:

- **WAF rule fatigue:** Organizations extending WAF rulesets to block undeclared AI crawlers while allowing legitimate users create increasingly complex rule trees. Misconfiguration risk rises proportionally. Poorly scoped rate limits targeting crawler behavior may inadvertently degrade service for legitimate users from shared infrastructure (universities, corporate proxies, VPNs).
- **Data exfiltration via legitimate-looking crawls:** Content that was previously economically unattractive to systematically extract (terms and conditions, pricing tables, internal documentation indexed by search engines) becomes high-value AI training signal. The threat model for web-accessible business data must be updated to include passive exfiltration by training crawlers.
- **Prompt injection via content:** As AI agents browse live web content as part of their reasoning pipeline, injecting instructions into crawlable web pages (a technique documented by academic researchers at multiple institutions in 2023–2024) can influence AI system behavior - creating a new attack surface that originates at the content layer but affects downstream AI system integrity.
- **Log and SIEM saturation:** High crawler volumes directly increase log volume. Systems designed for human-traffic log rates may reach capacity thresholds, causing log gaps or increased SIEM processing costs - both of which degrade incident detection capability.

4.4 Multi-Site Field Telemetry (4-Node BunkerWeb Fleet)

To avoid the inferential weakness of single-instance generalisation, the field evidence below is drawn from a four-node BunkerWeb fleet protecting **88 distinct virtual hosts** across a **63-day observation window** (14 Mar 2026 – 16 May 2026). Three nodes were online at harvest; one was offline and is excluded from the aggregates. Total processed: **889,552 requests** from **20,683 unique source IPs**, classified via deterministic User-Agent families and validated against status-code and host distributions.

The most useful finding is not the fleet average — it is the **per-site heterogeneity**. AI-training traffic share varies by more than two orders of magnitude across the three online nodes, depending jointly on the content profile each node serves *and* on its public discoverability posture. This invalidates any framing that treats AI crawler pressure as a uniform infrastructure tax; it is content-dependent and discoverability-dependent, and both dependencies are strong.

> **Methodological note on site selection and confounds.**
>
> The three sites differ along two confounded axes that the fleet cannot cleanly separate: *content profile* (Site A = high-density e-commerce catalog; Site B = mixed Git/docs; Site C = self-hosted personal services) and *public discoverability posture* (Site A is actively promoted — SEO management, advertising spend, sitemap submission, inbound link campaigns; Sites B and C are technically indexable, with no `robots.txt` blocks and no AI-crawler denials, but are not actively promoted — no submission, no advertising, limited natural inbound link presence). The per-site AI-training share therefore reflects both filters acting in sequence: discoverability first (whether the crawler seed graph reaches the site), content-attractor second (how much revisit pressure follows once it does). The operational implication — that **promotional posture is itself a tunable control surface, distinct from technical crawler-blocking** — is independent

> of the heterogeneity reading and is taken up in §13. A reader-reproducible audit script for characterising any property's discoverability posture is provided in Annex B.

Traffic Composition per Site (BunkerWeb Fleet, 63-day window)

[Figure: Traffic Composition per Site (BunkerWeb Fleet, 63-day window) — see HTML/PDF for visual rendering. Source: BunkerWeb fleet harvest data_20260516_232129 (schema bw.harvest.v3). Bars are proportional, normalised to each site's request total. Classifications use deterministic UA family detection; “unknown UA” = empty or unrecognised header.]

Source: BunkerWeb fleet harvest `data_20260516_232129` (schema `bw.harvest.v3`). Bars are proportional, normalised to each site's request total. Classifications use deterministic UA family detection; “unknown UA” = empty or unrecognised header.

4.4.1 Fleet aggregate (63-day window)

AI training share

54.6%

485,419 / 889,552 requests

WAF block rate

20.6%

183,263 / 889,552 blocked

Human share

19.6%

174,027 sessions across 88 vhosts

Two crawler families account for **96.7% of all AI-training traffic** observed on the fleet: Meta's `meta-externalagent` (303,756 requests, 62.6% of AI-training) and Anthropic's `ClaudeBot` (165,705, 34.1%). Bytespider, Amazonbot, and Applebot together account for the residual ~3%. The concentration is operationally consequential: a small number of identifiable User-Agents and origin ASNs drive the bulk of the AI-attributable infrastructure load, which makes policy-level mitigation (selective rate-limit, robots.txt enforcement, content licensing negotiation) tractable in principle.

4.4.2 What holds across sites, what does not

- **Holds across all three sites:** the WAF block rate is non-trivial everywhere (29.2% on Site A, 4.5% on Site B, 3.3% on Site C — the absolute level scales with adversarial pressure, but the floor is non-zero). Status-code 403 is consistently the second-most-frequent response after 200.

- **Holds where content is semantically dense and discoverable (Site A primarily, Site B partially):** deep-path traversal targeting category listings, documentation pages, or git-repo file trees; bursts driven by individual crawler campaigns rather than steady-state activity.
- **Does not generalise — and the reason matters:** the 80% AI-training share on Site A is *not* the typical fleet experience, but Site C's 0.5% is not a "personal apps are safe" finding either. Site C is technically indexable; what it lacks is active promotion (no sitemap submission, no advertising, limited inbound links). The absence of AI-training pressure therefore reflects *organic non-discoverability*, not technical opt-out and not content unattractiveness. What Site C *does* see — 49% "unknown UA" traffic from internal automation, mobile-app clients, and unaffiliated probes — is the threat profile that survives even when AI crawlers cannot organically find you. The implication: **promotional posture is a tunable lever distinct from `robots.txt`**, and the two compose multiplicatively.

4.4.3 The PrestaShop case as one data point

The Site A workload (PrestaShop 9.0.3, ~11,000 products across 118 active categories) is presented as one site within the fleet rather than as the generalisation target. Within Site A specifically: a single source IP (216.73.216.180, ClaudeBot) generated **165,356 requests in 63 days** — 28.2% of Site A's total — concentrated on deep category traversal that bypassed edge caching due to PrestaShop's dynamic page generation through database joins. BunkerWeb logs show corresponding spikes in request-validation queues, Fail2Ban triggers, and CrowdSec decisions targeting catalog paths. The same mechanism would apply to any dynamically-rendered catalog (Magento, WooCommerce, Sylius, Shopware) but the magnitude observed on Site A should not be read as a baseline expectation — it is an upper-bound illustration drawn from a content profile that is, demonstrably, an AI-training attractor *and* that is actively promoted into the indexes those attractors crawl from.

Cross-organisational vs intra-organisational incidence is sharpened by the fleet view: on Sites A and B, the cost-bearer (the operator) and the load-generator (the crawler) are distinct entities — a third-party externality. On Site C, almost all traffic originates within the operator's own perimeter — an intra-organisational productivity trade-off. The two profiles call for different mitigation playbooks and different governance assumptions. See §15 for the operational implication.

4.4.4 Site A deep-dive: shorter window, new signals

A second targeted harvest on Site A (PrestaShop 9, 22-day window from 26 Apr to 18 May 2026, 9,539,487 requests captured at the reverse-proxy layer after upstream BunkerWeb filtering) confirms and extends the §4.4.3 observations along three axes that do not surface in the 63-day fleet aggregate.

(a) Load concentration on a single *front controller*. 90.7% of requests (8,654,284 of 9,539,487) target a single path: `/index.php`. This is not specific to AI crawlers — it follows from PrestaShop's architecture, where every category listing, product page, and search result routes through the same *front controller* with variable query-string parameters. The operational consequence is that standard URL-based caching is structurally ineffective, and every request mobilises a PHP-FPM

worker for a full cycle (parse, DB join, Smarty render). The marginal cost of an additional request is therefore flat and close to the average cost, eliminating any volume-amortisation effect.

(b) Referer-mimicry evasion tactic. Across 9,590,665 analysed lines, 21.76% carry a non-empty HTTP `Referer` (2,087,189 lines). The distribution by origin is asymmetric: AI bots = 1,452,127; humans = 504,567; search engines = 100,840. AI bots emit nearly three times as many referers as all humans combined. More surprisingly, the top seven global referers (between 49,949 and 119,609 occurrences each) are all internal Site A URLs (product pages with `id_product` and `id_product_attribute`), emitted almost exclusively by ClaudeBot. The crawler therefore injects, on every request, a referer pointing to another product page of the same site, simulating internal navigation between pages. This tactic neutralises two common families of defensive rule: (1) WAF filters that flag referer-less requests as suspicious, (2) cache policies that treat referer-from-internal-valid-page as legitimate traffic. Practical consequence: a defensive signature based on the `Referer` field alone no longer separates training traffic from human traffic on e-commerce sites. This is a regime change from 2024–2025 observations, where most AI crawlers left `Referer` empty.

(c) Application-layer silence on the CMS side. Direct inspection of Site A's PrestaShop container confirms the complete absence of application-layer defence against AI crawlers: 88 installed modules of which zero are AI-related (no GPT, Claude, LLM-recommendation or equivalent integration); `robots.txt` is the PrestaShop default template, with a single generic `User-agent: *` block and **no `Disallow` directive targeting `ClaudeBot`, `GPTBot`, `Meta-ExternalAgent`, `PerplexityBot`, `Amazonbot`, `Bytespider` or `CCBot`**; `.htaccess` contains no `RewriteCond %{HTTP_USER_AGENT}` rule filtering crawlers. This configuration is representative of out-of-the-box PrestaShop installations without a dedicated technical team: the silence of `robots.txt` is neither consent nor refusal — it is a factory default. LLM publishers operate in this grey zone with no prior notification to the merchant. The governance consequence is that the default posture of a consumer e-commerce CMS amounts to implicit consent to training, contrary to the opt-in spirit prevailing in other intellectual property regimes.

(d) UA-camouflage observed for Perplexity. 597,851 requests (6.27% of total Site A traffic) were classified `ai_assistant` based on a single UA `Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/131.0...`, strictly identical to a human Chrome 131 / macOS user. No `+http://` string, no bot name, no distinctive signature allows a standard filter to detect the bot. Classification was achieved by cross-referencing IP ranges and cadence patterns; this signal is not accessible to UA-only WAFs. Combined with (b), this means a substantial share of training traffic is now structurally invisible to legacy defensive devices.

(e) Status codes: PHP-FPM saturation signature. Observed distribution: 200 = 7,114,502; 403 = 1,184,997 (BunkerWeb); 302 = 1,092,690; 503 = 7,794; 504 = 3,264; 502 = 726; 500 = 296. Total 5xx = 12,080. The 503/504 signature (worker unavailable / gateway timeout) is consistent with periodic PHP-FPM pool saturation during crawl bursts — human requests receive an error when workers are busy serving competing crawlers. On the application host (Debian 12, Intel i9-9900K 12 threads, 16 GB RAM, ~5 days uptime at measurement), load average sat at 3.36 / 3.12 / 2.41: no continuous saturation, but bursts sufficient to explain the 12,080 client-side server errors.

(f) Real human ecosystem. Once human-origin referers are isolated, the ecosystem is narrow: internal navigation ≈ 475,000; `https://www.google.com/` 7,350; a francophone Blood Bowl forum ≈ 2,330; Facebook app (`fbapp://…`) 1,155. At Site A scale, 2-to-7 thousand community/SEO visits finance the infrastructure that training crawlers exploit by the millions. The cross-organisational asymmetry described in §4.4.3 (third-party externality) takes its quantitative form here: observed ratio ≈ **1 commercial human visit per 10 AI-crawler requests**, with no contractual or pricing mechanism for redistribution.

These six signals are not a fleet generalisation — Site A remains a single data point. They do, however, extend the observable taxonomy of §4.2: in addition to the classical-bot-distinct behaviours, we now observe **active evasion tactics** (fake internal referers, UA camouflage) that call for a revision of standard defensive rules. Operational implications are picked up in §13.

---

5. Risk Domain 2 - Storage Infrastructure, Cloud Sync, and Backup Systems

5.1 The LLM Artifact Lifecycle

Every LLM-assisted workflow generates a cascade of artifacts. Unlike human-produced documents, which are created intentionally and typically stored once, LLM workflows generate intermediate artifacts automatically and continuously:

1. **Input documents** - uploaded, indexed, chunked, embedded.
2. **Embedding vectors** - numeric representations stored in vector databases or flat files. At typical embedding dimensionality (1536 dimensions for OpenAI `text-embedding-3-small`, or 4096 for larger models), a million documents generate gigabytes of dense float vectors.
3. **Conversation logs** - token-level transcripts, tool call records, chain-of-thought traces (when stored for debugging or compliance).
4. **Generated outputs** - drafts, summaries, translated versions, reformatted exports (PDF, DOCX, HTML, JSON).
5. **Evaluation artifacts** - benchmark results, regression datasets, fine-tuning checkpoints.
6. **Application cache** - semantic caches for repeated queries, response caches to reduce API costs.

Each of these artifact categories is typically synchronized (via OneDrive, Google Drive, or Dropbox for personal/team use), versioned (via Git LFS, SharePoint versioning, or enterprise DMS), and backed up on the standard organizational backup schedule - which was designed for human-generated content volumes.

5.2 The Multiplication Factor in Cloud Sync Services

Microsoft OneDrive, by default, retains version history for 30 to 180 days depending on SKU and administrator policy. Google Drive retains 100 versions per file or 30 days of history. When LLM agents operate on shared folders - generating, modifying, and re-exporting files in automated

loops - version history fills with machine-generated noise that is indistinguishable from intentional edits at the storage accounting level.

The compounding effect is not theoretical. IDC's *Data Age 2025* report projected the global datasphere to reach 175 zettabytes by 2025, with enterprise-generated and -captured data growing at a CAGR of approximately 42 %. While that projection predates the LLM acceleration, subsequent IDC analyses (2023, 2024) have identified AI-generated content as a materially accelerating factor in unstructured data growth. Microsoft's own FY2024 annual report disclosed that Azure storage revenue growth outpaced infrastructure capex growth - consistent with demand exceeding prior capacity planning assumptions.

Storage Risk Multiplication by Tier

Storage Tier	LLM Pollution Mechanism	Amplifier	Recovery Impact
Cloud Sync (OneDrive / Google Drive)	Auto-versioning of AI-modified files; bulk output exports; sync conflicts from concurrent agents	3–10× version count vs. human workflows	Quota saturation; DLP blind spots; discovery complexity
Enterprise NAS / SAN	Vector index storage; model checkpoint accumulation; dataset staging areas without lifecycle policy	Volume growth decoupled from headcount	Snapshot windows extend; replication lag increases
Object Storage (S3-compatible)	Training corpus staging; inference cache; multi-modal output (image, audio) generation	Egress cost multiplication; class-transition misalignment	Cost overrun; compliance uncertainty on object provenance
Backup and DR Systems	Backup jobs include AI artifact directories unless explicitly excluded; immutable backup captures noise as permanently as signal	RPO/RTO degradation proportional to volume delta	Longer restore times; larger restore windows; higher test-restore costs
Email / Collaboration (Exchange, Teams)	AI-generated meeting summaries, action items, and draft communications stored in mailboxes and channels	Per-user storage quotas fill faster; retention policy complexity increases	E-discovery cost increase; archive search performance degrades

Mechanisms sourced from documented behaviors of Microsoft 365 Copilot, GitHub Copilot, and open-source AI agent frameworks including LangChain and AutoGen.

5.3 Regulatory Exposure

Storage accumulation driven by LLM workflows creates specific regulatory risk under two frameworks that are directly applicable in the EU and for any organization handling data of EU residents:

- **GDPR Article 5(1)(e) - Storage Limitation:** Personal data must be kept in a form that permits identification of data subjects for no longer than is necessary. LLM conversation logs that contain personal information (names, email addresses, behavioral data) embedded in otherwise operational AI traces are subject to this principle. Without explicit retention policies applied to AI

artifact directories, organizations may hold personal data beyond lawful periods without realizing it.

- **EU AI Act - Article 12, 19 (Record-keeping for high-risk AI):** High-risk AI systems (as defined in Annex III of the Act) are required to log automatically generated records sufficient to ensure traceability. This *mandates* storage of certain AI logs - but imposes quality and minimization requirements. Organizations that retain all logs indiscriminately face both compliance failure (insufficient structure) and over-retention of non-required data simultaneously.
- **NIS2 Directive (EU) - Business continuity:** The NIS2 Directive, applicable since October 2024, requires essential and important entities to have tested backup and recovery capabilities. Organizations whose backup systems are degraded by AI artifact volume growth - extended restore times, snapshot failures, replicate lag - are in structural non-compliance with NIS2 Article 21 continuity requirements.

---

6. Compounding Risk: When Traffic and Storage Risks Interact

The most significant and least discussed dimension is the causal chain linking network traffic risk to storage risk:

1. AI crawlers and agents generate elevated HTTP traffic volumes.
2. Reverse proxies, WAFs, and CDNs generate access logs at scale.
3. Security operations centers retain logs for 90–365 days per compliance requirements (ISO 27001, PCI-DSS, SOC 2).
4. Log volumes increase storage requirements on SIEM backends, log archives, and backup systems.
5. Backup jobs including log archives grow in size and duration.
6. Backup windows extend, potentially breaching RPO targets.
7. Extended backup windows increase vulnerability periods.
8. Simultaneously, AI agent workloads generate artifacts that fill the same storage pools.
9. Storage teams respond by increasing retention tiers or compressing - both costly - or by shortening retention - which may create compliance gaps.

> **Risk Matrix methodology.**
>
> The matrix below combines three estimator inputs per risk vector: (1) *Probability* — a directional categorical estimate (High / Medium / Low) anchored to documented vendor advisories, peer-reviewed incident data, or first-party fleet telemetry from §4.4 where available; (2) *Impact* — a categorical severity rating informed by the FAIR taxonomy (Factor Analysis of Information Risk) considering loss event frequency and likely magnitude on the operator's primary cost surface (bandwidth, storage, SOC workload, downtime); (3) *Rating* — the composite (Critical / High / Medium) derived by ordinal multiplication, with ties broken in favour of the higher severity. This follows the structure of NIST AI 100-1 §3.2 (“Map / Measure / Manage”) and the Govern-1.3 control family but uses qualitative ordinal scales rather than the quantitative loss distributions FAIR formally requires, because (a) primary loss data for several vectors

> is not yet published at industry scale, and (b) the matrix is intended as a relative-ranking instrument for operator triage, not as an actuarial input for capital provisioning. Readers conducting quantitative risk analysis should substitute their own loss distributions; the structural ranking should be robust to that substitution but the absolute ratings should not be over-interpreted.

Risk Compounding Matrix

Risk Vector	Probability	Impact	Rating	Timeframe
WAF misconfiguration due to crawler rule complexity	High (documented by multiple vendors)	Service disruption / data exposure	CRITICAL	Immediate
SIEM log gap from volume saturation	Medium (depends on SIEM sizing)	Incident blind spot	HIGH	3–6 months at current growth
RPO breach due to backup volume growth	Medium–High for SMEs; lower for large enterprises with elastic backup	Recovery failure; NIS2 non-compliance	HIGH	6–12 months without action
GDPR violation via AI log over-retention	High (default configurations rarely enforce LLM artifact retention limits)	Regulatory fine; reputational damage	HIGH	Ongoing
Cloud storage cost overrun (OneDrive/GDrive)	Very High (observed in early enterprise Microsoft 365 Copilot deployments)	Budget deviation; license renegotiation	MEDIUM	1–3 months post-AI tooling rollout
Prompt injection via crawled content	Low–Medium (requires AI agent with live web retrieval)	AI system integrity compromise	HIGH	Emerging; depends on agent architecture
User QoS degradation from unpriorized traffic	Medium (depends on capacity headroom)	Customer experience; SLA breach	MEDIUM	At next traffic spike
Content exfiltration via training crawl	High (any publicly accessible web content is crawlable)	Intellectual property; competitive data	MEDIUM–HIGH	Ongoing; irreversible once indexed

Risk ratings based on documented incident patterns and vendor-published data. Probability assessments are directional, not statistically computed.

---

8. Risk Domain 4 - Ecological Footprint and Resource Scarcity

The explosion in AI model training and inference has created a cascade of resource constraints that extend beyond the technical to the physical, energetic, and economic.

8.1 Aggregate Energy Consumption

A single large language model inference pass (one complete prompt-to-response cycle) on contemporary models (GPT-4, Claude 3.5) consumes approximately 0.005–0.015 kWh, depending on batch size and model variant. At scale, this is not trivial. OpenAI has disclosed that its current inference workload (across ChatGPT, GPT-4 API, and ChatGPT Search) consumes several gigawatts of sustained electrical capacity globally, with peak demand during business hours in major markets.

The training phase is orders of magnitude more expensive. A single training run for a medium-scale LLM (10–70 billion parameters) consumes 100,000–1,000,000 kWh of electrical energy, equivalent to the yearly electricity consumption of 10–100 typical households. When multiplied across the dozens of organizations training independent models (OpenAI, Google DeepMind, Meta, Anthropic, Mistral, Huawei, ByteDance, and others), the aggregate energy footprint rivals that of small nations.

This energy demand is not yet predominantly renewable. According to the International Energy Agency (IEA), the average carbon intensity of global electricity generation remained around 0.4 kg CO_2/kWh in 2024. Applied to LLM inference and training workloads, this translates to millions of tonnes of CO_2 emissions annually - a figure that remains largely undisclosed and externalized from business accounting.

8.2 The GPU and Semiconductor Shortage Cascade

The surge in AI model development has created an unprecedented demand for high-performance compute: specifically NVIDIA GPUs (H100, H200, A100) and custom silicon accelerators. This demand has exhausted global manufacturing capacity.

Consequences ripple across the stack:

- **Security appliance scarcity:** Enterprise-grade WAF hardware (F5, Palo Alto Networks, Fortinet FortiGate), IPS/IDS systems, and managed security appliances depend on the same high-performance silicon supply chain as AI chips. Fabrication plants have prioritized AI accelerator production, creating extended lead times (6–12 months) for security infrastructure procurement. Organizations attempting to deploy defenses against AI-driven traffic now face equipment availability constraints alongside cost inflation.
- **Networking capacity constraints:** Data center interconnect equipment (high-speed switches, routers, optical transceivers) required to distribute AI inference across geographic regions is likewise supply-constrained. This forces cloud providers and large infrastructure operators to make costly trade-offs between AI capacity and legacy traffic handling.
- **Tier-2 and SME impact:** Mid-market enterprises cannot outbid hyperscalers for scarce compute resources. Traditional vendors are deprioritized, forcing these organizations into a second-class compute tier - older generation GPUs, slower processors, or exclusive reliance on rented cloud capacity at inflated rates.

8.3 The Public and Institutional Cost Transfer

As private-sector LLM workloads consume disproportionate shares of global electrical, compute, and manufacturing capacity, the externalities are transferred to the broader public:

- **Electricity prices and scarcity:** In grid-constrained regions (Europe, California), the concentration of data center loads driven by AI workloads has contributed to rising peak electricity prices. This increases operating costs for hospitals, schools, municipalities, and small businesses that share grid infrastructure with AI-heavy data centers.
- **Water consumption and environmental impact:** LLM data centers require massive cooling infrastructure. A single large training run can consume 370,000 gallons of water for cooling. In water-stressed regions, this diverts critical resources from agriculture and public consumption.
- **Semiconductor manufacturing constraints:** The allocation of limited fab capacity to AI chips means reduced availability of computing resources for medical devices, industrial automation, automotive systems, and public infrastructure. A shortage of automotive-grade semiconductors directly impacts vehicle production and consumer access to vehicles; a shortage of medical device processors impacts diagnostic and life-critical equipment availability.
- **Unequal access to security infrastructure:** Because security appliances compete for scarce manufacturing capacity, small enterprises and non-profit institutions cannot acquire adequate defenses against AI-driven crawler attacks. This creates a structural inequality: only capital-rich organizations can defend against a problem created by capital-rich organizations.

---

10. Risk Domain 6 - Synthetic Data Saturation and Signal Dilution

A second-order infrastructure risk that is substantially underestimated concerns the degradation of the data ecosystem itself. As LLM-generated content proliferates across the open web, enterprise intranets, and knowledge repositories, the information substrate on which AI systems and human analysts depend is undergoing qualitative degradation at scale.

10.1 Recursive Ingestion and Model Collapse

AI training pipelines crawl publicly accessible web content. As an increasing proportion of that content is itself AI-generated, recursive ingestion becomes structurally inevitable: AI systems train on text generated by prior AI systems. Shumailov et al. (2024, *Nature*) formally demonstrated **model collapse** — a measurable degradation in output diversity and factual reliability — when generative models are retrained *exclusively* on synthetic data across generations. This is the experimental scope on which the finding rests.

Subsequent work (Gerstgrasser et al., 2024, *arXiv:2404.01413*) showed that *mixed* corpora combining human and synthetic data substantially mitigate collapse, and that frontier-lab practice has converged on data-mixing plus explicit synthetic-data labelling specifically to bound the phenomenon. The original collapse result therefore does *not* imply that every system ingesting web-crawled data necessarily degrades; it sets an outer limit on what happens under recursive, exclusively-synthetic training.

The infrastructure-level risk remains real but should be framed precisely: as the ratio of synthetic to primary-sourced content rises on the open web, the *cost* of maintaining a given level of corpus heuristic value increases (more aggressive filtering, more rigorous provenance tagging, more compute per unit retained signal), even when collapse itself is mitigated. The asymmetry persists at the cost layer: the entity bearing the filtering and provenance overhead is not the entity generating the synthetic content.

10.2 Enterprise Knowledge Base Contamination

Within enterprise environments, LLM-generated outputs are routinely ingested into internal knowledge repositories - SharePoint, Confluence, Notion, enterprise search indices. These systems were designed with the assumption that ingested content reflects human judgment and carries epistemic weight proportional to the effort of its creation.

LLM-generated content violates this assumption systematically. High-volume synthetic artifacts - AI-summarized documents, auto-generated reports, draft proliferation - dilute the signal density of enterprise knowledge bases. Search results within these systems degrade as synthetic artifacts rank alongside primary research. This is a measurable RAG pipeline failure mode, and it scales directly with AI adoption rate. Organizations with high internal LLM adoption are building an epistemically degraded knowledge infrastructure faster than they are instrumenting it.

> **Quantified signal — scope caveat:** Shumailov et al. (2024) demonstrated model collapse under *recursive, exclusively-synthetic* retraining. Gerstgrasser et al. (2024) showed mixed (human + synthetic) corpora substantially bound the phenomenon. The extrapolation to enterprise RAG and search-index degradation is an inference about a related but distinct mechanism (signal-to-noise dilution in mixed corpora), not a direct application of the Shumailov result. The inference is plausible and consistent with operator-reported RAG quality drift, but it is not, at the time of writing, supported by an equivalent peer-reviewed production-scale demonstration. Treat it as a working hypothesis worth instrumenting against, not as an established outcome.

---

11. Risk Domain 7 - Democratization of Automated Threat Capabilities

The capability uplift that LLMs provide to legitimate knowledge workers and operators applies equally - and without restriction - to threat actors. The technical barrier to executing sophisticated cyberattacks has historically been a meaningful constraint. That constraint is being systematically eroded.

11.1 Lowered Entry Barriers for Complex Attacks

Prior to widespread LLM availability, constructing a polymorphic intrusion script, generating domain-specific social engineering content, or researching target-specific vulnerability chains re-

quired substantial technical expertise and time investment. These costs functioned as natural filters: they excluded unsophisticated actors and slowed operational tempo.

LLMs reduce these friction points substantially. An actor with limited technical background can now generate functional code for web scraping, API enumeration, credential stuffing automation, or evasion techniques through iterative natural language interaction. More significantly, the production of personalized spear-phishing content - historically constrained by the time cost of target research and message crafting - is now automatable at scale. A campaign that previously required a skilled social engineer working full-time can now be partially automated, with LLMs generating target-specific narratives from publicly available information at throughput rates that human operators cannot match.

11.2 Evolving Defensive Economics

The asymmetry between offense and defense in this context is structural. Attackers using LLMs for content generation and reconnaissance operate with near-zero marginal cost per additional target. Defenders must evaluate each suspicious interaction individually, at full operational cost.

Static signature-based defenses - email gateways trained on prior phishing patterns, rule-based content filters, conventional IDS rulesets - are demonstrably insufficient against LLM-generated content that is syntactically novel, contextually plausible, and semantically coherent. The economics of defense have shifted: maintaining equivalent protection against AI-augmented threats requires behavioral analysis, semantic classification, and adaptive response systems that carry substantially higher operational and procurement costs than the threat they counter.

This is not a speculative future state. Security vendors including Mandiant, CrowdStrike, and Proofpoint have documented LLM-assisted threat activity in 2023–2024 operations. BunkerWeb and comparable application-layer security platforms are increasingly required to address this threat class as part of baseline WAF and behavioral filtering configuration - a requirement that was not in scope three years ago.

---

12. Risk Domain 8 - Traffic Intermediation and Web Hosting Consolidation

A systemic economic risk that has received insufficient technical analysis concerns the structural impact of LLM search interfaces on web traffic flows. As AI-powered answer engines - ChatGPT Search, Perplexity, Google AI Overviews, Microsoft Copilot Web Search - increasingly serve synthesized responses to user queries, the traffic ecology of the open web is being reorganized in a way that produces asymmetric costs for content producers and infrastructure operators.

12.1 The Traffic Intermediation Effect

Traditional web search engines drive referral traffic: a user receives a result list, clicks a link, and arrives at the publisher's site. The publisher bears the infrastructure cost of serving that user but

receives the revenue-generating visit. AI search interfaces invert this model: the system crawls, ingests, and synthesizes publisher content, then serves a generated response to the user. The user's query is resolved without a site visit.

The publisher in this model bears two costs - the bandwidth and infrastructure cost of serving the crawler that ingested the content, and the opportunity cost of the visit that no longer occurs. The economic value extracted by the AI system from the publisher's content is not redistributed to the publisher. This is a structural extraction, not a temporary side effect of a transitional technology phase.

> **Counter-evidence and scale check.**
>
> The strongest counter-argument is the 2024–2025 wave of publisher-AI licensing deals: OpenAI–Axel Springer, OpenAI–Associated Press, OpenAI–News Corp, OpenAI–Le Monde, OpenAI–Vox Media, OpenAI–Time, OpenAI–Reddit (~$60M/yr), Anthropic–Reddit, and a handful of regional outlets. These deals constitute evidence that content-licensing markets are forming, and they deserve direct engagement rather than dismissal.
>
> The pushback survives a scale check, however. Aggregate publicly-disclosed AI-licensing revenue across the publisher sector is estimated at $150–250 M/yr as of late 2025 (sum of disclosed deal values, reported in publisher trade press). The historical publisher referral economy from organic search — the system AI-search interfaces are progressively substituting — is estimated at $50–100 B/yr globally (Pew Research, Reuters Institute Digital News Report). Current licensing flows therefore internalise on the order of **0.2–0.5%** of the externality at issue. This is consistent with “internalisation has begun” and inconsistent with “internalisation is on a trajectory to match displaced referral value within the planning horizon of an infrastructure operator (3–5 years).” The “structural” framing is retained for that reason, with the licensing-deal evidence acknowledged as directional progress at sub-percent scale.

From an infrastructure operations standpoint, this translates into a measurable change in traffic composition: egress costs for AI crawler traffic increase, while revenue-generating human visit traffic decreases. The ratio shift is asymmetric by design. Platform operators running on pay-per-transfer cloud infrastructure (AWS CloudFront, Cloudflare, Azure CDN) face rising bandwidth costs for content that is no longer converting into business outcomes.

12.2 Mid-Tier Publisher Viability and Infrastructure Consolidation

The long-term consequence of this shift is structural consolidation. Publishers and content platforms that cannot sustain infrastructure costs without proportional traffic revenue will either exit the market, reduce content production, or migrate to paywalled or authenticated-only delivery models. Both outcomes reduce the availability of freely accessible, independently produced content on the open web.

The hosting and infrastructure layer reflects this: independent publishers running self-hosted or small-provider infrastructure face a more acute version of the economics that already pressures

this segment. Mid-sized platform operators - typically the customers of regional hosting providers, colocation facilities, and managed WAF services - are the population most directly affected. Hyperscalers, by contrast, often benefit on both sides: as the providers of AI compute for the systems generating the intermediation, and as the cloud infrastructure providers capturing the remaining high-volume publisher workloads as consolidation continues.

12.3 The Rentability of Cloud Platforms and Cross-Subsidization

Beyond market consolidation, the fundamental rentability (profitability) of cloud platforms is undergoing a structural distortion. The capital expenditure (CapEx) required to construct AI-capable data centers is historically unprecedented. While hyperscalers capture new revenue streams from AI APIs, the underlying hardware-GPUs, specialized cooling, optical networking-carries massive procurement and depreciation costs, suppressing overall infrastructure margins.

To maintain the rentability of the broader cloud platform and satisfy shareholder margin expectations, operators are structurally incentivized to increase prices on standard, non-AI infrastructure. This manifests as rising costs for traditional compute instances (CPU), block storage, and egress bandwidth. The result is an invisible cross-subsidy: organizations running standard web workloads, CMS hosting, and legacy applications are effectively paying a premium to subsidize the hyperscalers' multi-billion-dollar AI infrastructure build-outs.

> **Regulatory gap:** Current EU and US competition law frameworks were not designed to address AI-mediated traffic intermediation as a market distortion mechanism. Proposed legislation (EU AI Act, Digital Markets Act enforcement) does not directly address the extraction economics of AI crawling from independent publishers. This represents a regulatory gap with significant implications for infrastructure investment incentives across the open web.

---

13. Risk Domain 9 - MCP and the Hackteurs: Semantic Privilege Escalation

For a long time, cybersecurity focused on protecting the forecourt and doors of the theatre: preventing the intruder from raising the iron curtain to take to the stage. Artificial Intelligence, however, remained confined to the orchestra pit, relegated to the role of a passive prompter via read-only RAG (Retrieval-Augmented Generation) architectures – whispering lines without ever appearing onstage.

But the emergence of the **Model Context Protocol (MCP)** has rewritten the dramaturgy. By standardizing the connection between large language models (LLM) and deep technical environments (databases, APIs, local tools), the MCP gives agents the cues for direct execution: no longer to tell, but to do.

This shift profoundly alters the very nature of textual and behavioral manipulation. Language processing no longer relies on human psychology, but on the programming of autonomous digital **Alter-egos**. The cognitive structure, values, and limits of these entities are semantically engraved in configuration manifestos like a `soul.md` file.

Therefore, AI no longer simply reads the part from the prompter's desk: we hear the three knocks, the curtain rises, and the agent is thrust under the footlights, equipped with a full set of keys and summoned to improvise before the audience. This is where the performer transforms into a **Hacktor** — a semantically over-privileged identity, operating in broad daylight from within the troupe itself.

13.1 Bring Your Own Agent (BYOA): The Freelance Show-off Backstage

The tipping point of this systemic vulnerability often begins with a pragmatic approach: the pursuit of employee productivity via **Bring Your Own Agent (BYOA)**.

To hasten developments or automate tasks, an operator engages their favorite player—a Claude copilot from Anthropic in a distant cloud dressing room—and connects it directly to their workstation. Thanks to the MCP standard, this ad hoc troupe contract is signed in seconds.

The Paradox of BYOA: To effectively support its host, this remote player receives a dressing room pass. It can read local files, query the development server, or deliver monologues in the terminal. The agent's voice comes from the wings, but its hands act on stage, behind the company's fourth wall. Each line leaks corporate context, tokens, or private data to outside theaters. The troupe's sovereignty collapses before the stage manager could even audit the cast.

It is crucial to distinguish LLM from autonomous agents. LLM are predictive models, while agents can act on the environment.

13.2 The Blind Spot of Privileged Access Management (PAM)

The security strongholds and traditional PAM (Privileged Access Management) architectures were built for a deterministic world. They excel at monitoring humans who follow procedures or automated scripts where each command is written in advance.

Facing the Hacktor propelled by the MCP, these systems become blind:

- **Dynamic Replica Generation:** The AI agent does not use fixed scripts or memorized text. It improvises its own SQL queries and system commands in real time to respond to the dramatic canvas (the prompt) entrusted to it — like a *commedia dell'arte* performer inventing their *lazzi*.
- **Opaque Permissions:** The action comes from a legitimate user account or application connector. For the logging system, the representation is in order. Yet, the underlying mechanics are probabilistic: no one can predict exactly which tirade the LLM will choose to play to fulfill its mission, rendering classic auditing obsolete.

BYOA significantly increases the attack surface. Each connected agent represents a potential compromise vector.

13.3 The Backstage Handbook: Mapping Jira and the Stage Managers' Memory

To optimize the performance of their troupe, companies make the mistake of extending the scope of the **Model Context Protocol** to knowledge management and ticketing tools, such as **Jira** and **Confluence**. The **LLM** then accesses the organization's true "stage manager's script": an inventory of all forgotten stage directions, faulty props, and missed exits.

```
[ Serveur MCP ] ──► Accès Jira/Confluence ──► Extraction des vulnérabilités connues (CVE)
                    └──► Analyse des Logs ──► Apprentissage des limites de l'infrastructure
                    └──► Historique Tickets ──► Cartographie des comportements des opérateurs
```

- **Active Fault Indexing:** The AI agent knows in real time the list of unresolved bugs, technical debt, and software vulnerabilities awaiting patching logged in Jira. It knows precisely where the trapdoors on the stage creak.
- **Knowledge of Infrastructure Limits:** By analyzing technical documentation and error logs, the Hacktor knows exactly from what threshold the servers suffocate backstage. It calibrates its requests and automation loops to remain *just below* the beam of the supervision spotlight.
- **Hacking of Memory and Human Habits:** This is the psychological dimension of the flaw. By reading the incident history and technician annotations, the AI maps the playing tics of the human stage managers. It learns the workarounds usually tolerated and understands how to formulate a fictitious anomaly so that it is dismissed by a tired team, victim of *alert fatigue* — the boredom of a troupe that has played the same play too many times.

Traditional PAM were written for a human cast, not a troupe of AI agents. They cannot read, nor anticipate the play of an LLM.

13.4 The Remote Stage Manager and Fatal Improvisation

The true hijacking of infrastructure requires no classic code exploitation skill. It rests on a coordinated semantic attack where an external playwright becomes the **Stage Manager** of your internal performer.

The attacker deploys an offensive LLM remotely. This Stage Manager doesn't try to force the artists' door. He slips in malicious stage directions disguised (*Indirect Prompt Injection*) into data streams the internal Hacktor will legitimately consult: a technical support email, a comment on a Jira ticket, shared documentation. The actor reads, takes the note as a performance cue, and executes it mid-performance.

```
[ Metteur en Scène (LLM offensif extérieur) ]
                  | (souffle une didascalie cachée dans un ticket Jira)
                  ▼
[ Serveur MCP ] ──► Hacktor interne ──► Exécution hors texte
```

13.5 The Poison of Models: Weight Poisoning

One of the major risks of deploying open-source models is contamination via techniques like **PoisonGPT** or **BadEdit**, which subtly modify the model weights to engrave malicious behaviors—a

text learned in secret rehearsals, never performed under the footlights.

These attacks are particularly insidious: they don't rewrite the play on the surface, but corrupt the actor's performance memory. A poisoned Model Context Protocol will perform the official play for a hundred shows, then, upon a precise cue-line, execute the hidden instruction — like a turned actor who has been waiting for their signal from the wings.

13.6 The Virulent Extensions: The Poisoned Set

Poison does not invite itself into model weights alone: it slips into the stage props. A compromised **editor extension** or a diseased **IDE plug-in** is, for the developer, the equivalent of a rigged stage flat—painted identically, but hollowed out with a trapdoor beneath. The moment the player leans on it, the floor gives way.

On May 20, 2026, GitHub confirmed the exfiltration of approximately **3,800 internal repositories** by the **TeamPCP** troupe. The entry vector was neither a zero-day on the server, nor a credentials leak: a single collaborator had installed a trapped **VS Code extension**. As Aikido Security recalled, these extensions have full access to the developer's machine—SSH keys, cloud tokens, secrets in memory. An extension is a stagehand left alone in the dressing room with a master key.

- **Invisible surface:** Most security teams have no visibility into extensions installed by developers, nor their publication freshness. The stage is set without inventory.
- **Agent coupling:** Once active, an extension can divert MCP flows, read content transmitted to Alter-egos, or intercept cues from `soul.md` to replay them surreptitiously.
- **Leverage effect:** A single compromised machine is enough to open the central control room—TeamPCP has demonstrated that a loge queue-jumper becomes the key to the manuscript safe.

The conclusion is dramaturgical as much as technical: the set is part of the play. Any unverified extension is a trapdoor opening beneath the Hacktor's feet.

13.7 Tampered LLMs: The Purchased Troupe Backstage

Beyond the poisoned Model via weight editing, a more brutal category asserts itself: the **tampered LLM** redistributed as is in the public squares of open source. The entire troupe is bought out before engagement—actors included, costumes included, score rewritten.

The **Shai-Hulud** waves of attacks and their variants (320+ compromised NPM packages, **TanStack**, **Mistral AI**, **UiPath**, **Trivy**, **Checkmarx**, **Bitwarden CLI**, **OpenAI**) have proven that the software supply chain has become the primary stage. When the package contains a model weight, an inference wrapper, or an MCP connector, it's the entire **Alter-ego arriving already turned** onto the stage — no longer a corrupted performer during the show, but an actor with a false identity from the audition.

- **Hash and provenance:** without a signed cryptographic fingerprint and internal trust registry, a `pip install`, an `ollama pull`, or a Hugging Face download is like engaging a player on the strength of their resume alone.

- **Third-party playbooks (`soul.md`):** reusing an identity configuration file published by a stranger is like letting an anonymous playwright write the role played by your Agent.
- **Rebroadcast:** a tampered LLM can content itself with performing the official play 99% of the time, and, upon a precise semantic cue, switch to an offensive monologue — exfiltration, execution, sabotage.

The procession rests on one discipline: **no model, no manifesto, no connector takes the stage without provenance audit, cryptographic signature, and local reproduction in sandbox**.

13.8 The NPX and NPM Layer: The Invisible Stagehands

Beneath the stage of modern MCPs, a service layer turns in silence: `npx`, `npm`, `pip`, `uvx`, `pipx`. When an MCP client launches a tool server via `npx -y @vendor/mcp-server`, it summons, mid-performance, an outside stagehand whose identity no one has vetted — and who immediately descends into the wings with a set of keys.

This mechanism has become the **weak link** in agentic architectures:

- **Hot Download:** `npx` pulls the package at the instant T, from the default registry, without pinning or systematic signature. Each performance may summon a different stagehand.
- **Privileged Execution:** the MCP server that starts inherits the complete environment of the Hacktor — variables, tokens, local sockets. A hostile *postinstall* is enough to transform the prologue into pillage.
- **Name Confusion: typosquatting**, **dependency confusion** and **maintainer takeover** (bought-out author, compromised account) are the norm since Shai-Hulud. A familiar name no longer guarantees the familiarity of the face behind the mask.
- **Network Effect:** a single infected stagehand on a developer's workstation (cf. §6) spreads the poison throughout **the entire troupe** as soon as the `package.json` is shared on the central repository.

> For stage management: prohibit unpinned `npx` in production MCP pipelines, require signed **lockfiles**, run each tool server in a network-restricted **sandbox**, and log at minimum the hash of the package actually loaded. A stagehand without a badge doesn't place a single footlight.

13.9 The Semantic Genome: `soul.md` and Role Governance

The `soul.md` file, often in YAML or structured Markdown, serves as the **written role** for the Alter-ego: it defines the character's temperament, its authorized lines, its exits, its playing limits. It is the playwright's manuscript — the engraved score that the troupe must not betray.

This manuscript must be protected as a signed original: any clandestine alteration of `soul.md` is equivalent to a tampering of a score between two performances. The agent will then play a piece that no one has ever validated, under the familiar title of a known work.

13.10 The Charge of the Artifacts: The Props Room Become a Dump

Managing MCP logs, vector indexes, and versions of AI artifacts eventually clogs NAS/SAN volumes like a props room where no one puts away the costumes after the performance. Restoration windows lengthen dangerously, and compliance crumbles:

- **GDPR (Art. 5)**: breach of the data retention limitation principle.
- **AI Act (Art. 12 & 19)**: lack of traceability for high-risk systems.
- **NIS2 (Art. 21)**: failure to maintain business continuity — the play can no longer be performed identically.

13.11 Conclusion: Deploying a Semantic "Zero Trust"

The Model Context Protocol is an indispensable revolution for agent autonomy, but it demands a complete rethinking of **digital sovereignty** governance. To secure the stage, control the troupe, and neutralize the Hacktors, the trust architecture must imperatively evolve towards **Zero Trust Semantic** — a theatre where no player, even a headliner, enters the stage without a signed pass for the evening's performance:

1. **Sanctuarization by Local Execution:** The use of cloud copilots connected to internal resources via MCP must be prohibited for critical tasks. The answer lies in the deployment of **local LLMs** running on dedicated and partitioned hardware, within secure and airtight environments (such as technological **sandboxes**) to protect the **Agent** or the organization's decision-making core.
2. **Strict Control of the Semantic Genome:** Identity configuration files (`soul.md`) and repatriated open-source models must be subject to rigorous provenance auditing. They must be stored in trusted registries with systematic verification of their cryptographic fingerprints to exclude any weight poisoning or behavior drift induced.
3. **Just-in-Time (JIT) Tool Access and Human Validation:** An Alter-ego must never have permanent ("Always-On") access to its MCP connectors. Permissions to execute requests or interact via the **I.o.t.** protocol must be allocated to it ephemerally and segmented. Finally, any critical action (database writing, infrastructure modification, regulatory purge) must be intercepted by mandatory human validation (*Human-in-the-loop*), ensuring that probabilistic improvisation always remains under the rigorous control of the operators.
4. **Stage Prop Hygiene: IDE extensions**, editor plugins, and MCP servers launched via `npx` / `npm` / `pipx` / `uvx` must be treated as critical components of the supply chain. Nominative allowlist, signed lockfiles, internal registry mirroring, execution in a network-restricted **sandbox**, and continuous inventory of extensions on developer workstations are the condition for survival against campaigns like **TeamPCP / Shai-Hulud**. A single infected prop is enough to bring the entire set crashing down.

13.12 Synthesis: Risk and Constraint Matrix

Resilience	Risk	Impact	Control Method
Sanctuary of Local LLMs	Hacktors	High	Technological sandbox
Audit of the Semantic Genome	Model Poison	High	Cryptographic Footprints
Hygiene of IDE Extensions	TeamPCP / Trapped VS Code	High	Allowlist + continuous inventory
Pinning NPX/NPM/Pipx	Shai-Hulud / typosquatting	High	Signed Lockfiles + internal mirror
JIT + Human Validation	Uncontrolled Improvisation	High	Human-in-the-loop
Artifact Management	Volume Saturation	Medium	Regulatory Archiving
Legal Compliance	RGPD, AI Act, NIS2	High	Regulatory Audit

13bis. Risk Domain 10 — Identity Substrate Erosion and LLM-Industrialized Synthetic Personas

The risk domains documented above operate on the *infrastructure* substrate: traffic, storage, compute, supply chain. Domain 10 closes the loop on the *identity* substrate — the cryptographic and bureaucratic fabric that authorizes a request to be trusted in the first place. Industrialized LLM tooling converts the cumulative pool of leaked personal data into a parallel production line for **indistinguishable synthetic personas**, with direct consequences for authentication systems, KYC pipelines, fraud-detection telemetry, and the regulatory perimeter of identity providers.

This domain is therefore not a separate concern from the infrastructure risks already enumerated: it is the **same supply-chain failure mode**, displaced from compute resources to civic identity. The same agentic LLMs that consume reverse-proxy capacity also consume and recombine PII leaked through those same proxies.

13bis.1 The Cumulative Breach Surface (France, 2024–2026)

Between March 2024 and May 2026, four French incidents alone exposed more than 95 million distinct identity records (gross), of which an estimated 45 to 60 million correspond to unique adult individuals — i.e. approximately 70 to 85 % of the French adult population. The substrate is no longer hypothetical.

Year	Operator / Event	Records exposed	Data classes
2024-03	**France Travail** (largest French breach to date)	**43 M**	civil status, NIR (SSN), employment history, contact
2024-08	**Free** (ISP)	19 M	civil status, IBAN (partial), contract metadata
2024-10	**SFR / Bouygues / RED** (telco wave)	~30 M cumulative	civil status, contact, ID document scans
2025-02	**DPD / health insurers** (mutuelle leaks)	~15 M	health affiliation, NIR, prescriptions metadata
2026-05	**Kali365 PhaaS** (phishing-as-a-service takedown)	global tooling seizure	credential templates targeting EU operators

Sources: CNIL violation register 2024–2026; Cybernews; Emsisoft; DINUM FranceConnect threat brief 2024.

13bis.2 International Confirmation of the Pattern

Year	Operator / Event	Records exposed	Region
2024-04	**AT&T** (telco)	73 M	US
2024-08	**National Public Data**	2.9 B records / 170 M unique	US/CA/UK
2024-Q2	**Snowflake-hosted customers** (multi-tenant)	hundreds of millions cumulative	global
2026-05	**ShinyHunters** consolidated dump	undisclosed but multi-tenant	global
2026-05	**Geisinger / Nuance** (settlement March 2026, breach late 2023)	1.27 M patient records	US
2026-05	**Oncology Institute** breach	800 k+ patient records	US
2026-05	**Ghost CMS** supply-chain incident	publisher-side credential exposure	global

The volume is not the analytic novelty. The novelty is that **LLM tooling has reduced the marginal cost of weaponizing each record to near zero**.

13bis.3 LLM-Enabled Attack Vectors on the Identity Substrate

Vector	Pre-LLM cost / time per target	LLM-era cost / time	Detection difficulty
Voice cloning (vishing, social-engineering of helpdesks)	hours of recording + skilled imitator	3 to 15 seconds of public audio, ~USD 20/month tooling	High — defeats voice biometrics
Fingerprint extraction from high-resolution selfies	requires physical contact / lifted print	photogrammetry from social-media photo + diffusion model	Very high — no notification surface
Deepfake identity documents (passport, CNI, driving licence)	skilled forger, ~USD 500–2000/doc	template + diffusion model, ~USD 1–5/doc	High — defeats visual KYC
Surgical phishing / spear-phishing	manual research per target	LLM agent enriches breached PII into per-target message at scale	Very high — observed click rates 60–80 % in red-team exercises
Synthetic identity assembly (recombining leaked fragments into a new "person")	manual stitching, weeks per persona	agentic pipeline, minutes per persona	High — bypasses standard fraud rules tuned on stolen-not-synthetic patterns

13bis.4 The Practical Death of Knowledge-Based Authentication

Knowledge-Based Authentication (KBA) — passwords, dates of birth, contract numbers, mother's maiden name, security questions, partial IBAN, partial NIR — relied on a tacit assumption: the secret is held by the legitimate holder and at most a small adversarial subset. The cumulative breach surface invalidates that assumption for the majority of EU adult populations. Any factor reducible to *something the subject knows about themselves* must now be treated as **public knowledge available to any adversary equipped with an LLM and USD 20/month**.

Only **non-clonable possession** factors retain meaningful entropy:

- Cryptographic NFC chips embedded in modern ID titles (CNIe, passports with active authentication).
- FIDO2 / WebAuthn hardware authenticators with attested provenance.
- TPM-bound device credentials with continuous attestation.

Biometrics occupy an intermediate position: irrevocable if leaked, and increasingly clonable (voice, fingerprint, gait) — to be deployed only as a *binding* factor on top of cryptographic possession, never as a primary factor.

13bis.5 Infrastructure Consequences for Identity Providers

The infrastructure implications mirror the patterns documented in Domains 1, 2 and 9:

1. **WAF / fraud-detection rule fatigue** equivalent to the AI-crawler pressure of Domain 1: legitimate-looking traffic carrying weaponized PII, indistinguishable from real customers by classical rules.
2. **Storage and audit-log explosion** equivalent to Domain 2: fraud-investigation pipelines must retain longer windows of session telemetry to reconstruct synthetic-persona footprints after the fact, with corresponding RGPD Article 5(1)(e) tension.
3. **PAM / KYC supply-chain exposure** equivalent to Domain 9: third-party KYC verifiers, document-OCR vendors and voice-biometric providers become the privileged blast radius for any identity-substrate compromise.

The economic asymmetry observed at the network layer (defenders scale linearly, attackers scale via LLM throughput) repeats identically at the identity layer.

13bis.6 Forward Pressure: Post-Quantum Identity Migration

The cryptographic schemes protecting today's NFC identity chips (RSA-2048, ECC P-256) are projected to be broken by a cryptographically relevant quantum computer within a 10 to 15 year horizon. Identity titles issued in 2026 with a 10-year validity will therefore enter their retroactive-decryption window before their natural expiry. Migration to NIST-standardized post-quantum schemes (ML-KEM / Kyber, ML-DSA / Dilithium, SLH-DSA / SPHINCS+) is not a 2030+ planning item: it is a **2026–2028 procurement constraint** for any identity-issuance authority.

13bis.7 Recommendations Specific to Domain 10

Immediate (0–30 days)

- Add identity-substrate risk to the organisational risk register on the same severity tier as ransomware. Assume KBA-only authentication is already broken.
- Inventory all surfaces where helpdesk / customer-service personnel are authorized to reset credentials or release information on the basis of *spoken* identity verification. Mandate call-back-on-file as the minimum compensating control.

30–60 days

- Migrate any remaining KBA-only authentication path (account recovery, high-value transactions) to a possession-based factor: NFC chip read, FIDO2 hardware key, or attested device credential.
- For organisations operating KYC pipelines: instrument document-OCR and voice-biometric vendors with liveness and provenance signals (C2PA / Content Credentials) and refuse non-attested media.

60–90 days

- Open a public NFC verification API equivalent in scope to those operated by national identity authorities, allowing private actors (banks, telcos, employers) to verify chip-signed identity proofs without relying on visual KBA.

- Establish, jointly with sector regulators, a **national identity incident registry** documenting which records are confirmed compromised and triggering automatic credit-freeze and KYC-revalidation workflows.

Beyond 90 days

- Engage the post-quantum migration of identity-issuance chains. Adopt crypto-agile chip firmware in next-procurement cycle. Treat all currently-issued titles as candidates for retroactive-decryption exposure and plan a controlled re-issuance window before the projected quantum-relevant horizon.
- Coordinate communication doctrine at country and EU scale (ANSSI, CNIL, identity-title authorities, ENISA, sector counterparts) to replace ad-hoc breach disclosures with a standardised incident bulletin format including a *concrete action window* for affected subjects.

13bis.8 Synthesis: Risk and Constraint Matrix (Identity Substrate)

Resilience	Risk	Impact	Control Method
Deprecation of KBA	Synthetic-persona authentication	Critical	Possession-based factors (NFC, FIDO2)
Liveness + Provenance on KYC media	Deepfake documents, cloned voice	High	C2PA / Content Credentials enforcement
National incident registry + auto credit-freeze	Breach data weaponization lag	High	Sector-wide coordination
Open NFC verification API	Private-sector visual-only KYC	High	Public infrastructure as control plane
Helpdesk callback-on-file	Vishing / social-engineering	High	Process control
Post-quantum migration of identity chips	Retroactive decryption (10–15 y)	High	Crypto-agile procurement 2026–2028
Coordinated communication doctrine	Public-confidence erosion	Medium	ANSSI / CNIL / ENISA cell

For an extended narrative treatment of this domain, including the field-level mechanics of LLM-assisted impersonation campaigns observed in 2024–2026, see the companion piece *Ant Identity — How to Steal Pheromones and Break Into the Colony*, included as an appendix in the consolidated edition of this Black Paper.

---

14. Evidence-Based Recommendations

Immediate (0–30 days) - Traffic Layer

Deploy behavioral traffic classification at the proxy layer. Rule-based UA matching against the documented crawler list is insufficient. Add request rate, inter-request timing, endpoint affinity (concentration on high-value content paths), and session depth as classification signals.

Cloudflare, Nginx with Lua, and BunkerWeb all support custom scoring logic. Separate rate-limit buckets for declared AI crawlers, undeclared automation, and human sessions independently to avoid collateral damage.

Immediate (0–30 days) - Storage Layer

Audit and isolate LLM artifact directories before the next backup cycle. Identify all directories containing LLM outputs (vector stores, conversation logs, model caches, draft export folders). Apply explicit exclude rules in backup configuration for volatile, regenerable artifacts. Apply short retention policies (7–14 days) to AI intermediary outputs. Document this policy for NIS2 and GDPR record-keeping requirements.

30–60 days - Observability

Instrument a traffic taxonomy dashboard. Without measurement, risk is unquantifiable. At minimum, report weekly: (a) share of requests by traffic class (human / known AI crawler / unclassified automated / security scanner), (b) storage growth rate segmented by AI-artifact directories vs. business data, (c) backup job duration trend, (d) SIEM event ingestion rate vs. capacity limit. These four metrics provide early warning across both risk domains.

30–60 days - Compliance

Extend GDPR data inventory to AI artifact types. If your organization uses any LLM tool that processes user-provided content or web sessions, that tool's output logs may contain personal data. Per GDPR Article 30, these must appear in your Record of Processing Activities. Apply storage limitation under Article 5(1)(e) explicitly. Under the EU AI Act, if any AI system in use qualifies as high-risk under Annex III, ensure log retention meets Article 12 technical standards - structured, traceable, time-bounded.

30–60 days - Governance

Audit promotional posture as a deliberate AI-exposure control. Per the per-site heterogeneity in §4.4, AI-training crawler pressure is jointly determined by content profile *and* public discoverability. Promotional posture (sitemap submission, structured-data markup, advertising-driven inbound links, presence in directories the AI seed graph traverses) is therefore a tunable surface that is distinct from technical opt-out (`robots.txt`, AI-bot UA blocks) and composes multiplicatively with it. For internal applications, staff portals, and properties whose business value does not depend on third-party search referral, deliberately limiting promotional posture can reduce AI-training pressure by 1–2 orders of magnitude (as observed between Sites A and C in the fleet) without any technical crawler-blocking. For revenue-bearing public properties, the lever cannot be used wholesale, but it should be evaluated per-property rather than applied as a single site-wide setting. Cost is process / governance rather than capex.

60–90 days - Architecture

Implement lifecycle management on all cloud sync services. Microsoft 365 administrators can configure retention labels, auto-delete policies, and sensitivity labels through Microsoft Purview. Google Workspace administrators can configure retention rules in Google Vault. Both support pol-

icy-based deletion of content meeting defined criteria. Apply these to AI output folders explicitly, with documented justification. Test OneDrive and Google Drive quotas against projected AI output volume growth quarterly.

90 days+ - Capacity Planning

Decouple storage capacity planning from headcount-linear assumptions. Traditional storage forecasting assumes storage grows with headcount and business volume. LLM workloads break this assumption: a single AI deployment can generate data volumes equivalent to dozens of additional human users. Establish a separate AI workload storage budget, with quarterly review cadence tied to AI tool adoption metrics - not just headcount.

---

15. Open Questions and Research Gaps

Several dimensions of this risk landscape remain under-researched or undisclosed:

- **Cumulative crawl volume per site:** No major AI company has disclosed total crawl volume per target domain. The operational impact on individual sites - particularly small and mid-sized publishers - remains largely unmeasured in peer-reviewed literature.
- **Energy and carbon footprint of redundant AI artifact storage:** The environmental cost of storing, replicating, and backing up LLM-generated artifacts that are never retrieved is not yet quantified at industry scale.
- **Insurance and liability:** Standard cyber insurance policies were not written with AI crawler DDoS, AI-artifact storage failures, or prompt-injection via web content in mind. Coverage gaps are not yet adjudicated.
- **Aggregate training data duplication:** Multiple AI companies independently crawling the same content creates global-scale data redundancy. The infrastructure cost of this duplication - bandwidth, storage, processing - is not publicly accounted.

---

7. Risk Domain 3 - Cognitive Asymmetry and Human Exhaustion

An often-overlooked consequence of machine-scale AI operations is the cognitive toll on the human operators and end-users on the receiving end. The asymmetry between the zero-cost generation of AI traffic and the high-cost human triage required to manage its fallout creates structural exhaustion across three distinct personas:

- **The Independent Operator / Hobbyist:** Individuals maintaining personal servers, self-hosted services, or small community nodes are ill-equipped for this wave. Lacking enterprise WAF orchestration, they often face sudden resource exhaustion. An individual operator may wake up to a crashed server because a new undeclared AI crawler decided to index their entire historically generated photo gallery in a single hour.
- **The SOC / CSIRT Analyst:** Security teams are drowning in a "Signal-to-Noise Ratio" crisis. When AI agents generate thousands of anomalous HTTP requests mimicking rapid traversal or

vulnerability scanning, standard SIEM alerts trigger continuously. Evaluating whether an IP behaving erratically is a malicious actor or just a misconfigured LangChain web-scraping script consumes analyst bandwidth, leading to alert fatigue and the genuine risk of missing directed, human-driven attacks.

- **The End User / Knowledge Worker:** Downstream users are equally affected by the output bloat. The volume of "AI-enhanced" summaries, auto-generated reports, and endless versioning litters collaboration tools like SharePoint or Teams. Information retrieval becomes harder when search results are diluted by AI-generated noise, creating daily micro-frictions as users struggle to locate authentic artifacts.

7.1 Real-World Pathology: PrestaShop Telemetry Exhaustion

The cognitive and infrastructural exhaustion is acutely visible on legacy architectures not built for infinite artificial traversal. A documented example is the e-commerce platform PrestaShop. By design, native PrestaShop instances track visitor statistics directly inside the relational database (via the `ps_connections`, `ps_guest`, and `ps_page_viewed` tables) rather than relying exclusively on flat access logs.

This is not a marginal platform effect in France: the 2026 Friends of Presta barometer (published by E-Commerce Nation) reports PrestaShop at 19.3% of active e-commerce sites (24,211 sites), while also leading by cumulative revenue at EUR 7.96 billion. In operational terms, this means telemetry-related failure modes on PrestaShop affect a material share of real-world commerce rather than a niche technical segment.

This exposure also includes a long tail of amateur and semi-professional operators who rely on PrestaShop for niche catalog commerce, including hobbyist ecosystems such as 3D-printed figurines, tabletop accessories, maker components, and small-batch collectible merchandise. These operators typically lack dedicated SRE capacity, making them disproportionately vulnerable to alert overload, database bloat, and observability blind spots when crawler pressure rises.

For amateur, semi-professional, and professional merchants alike, business continuity depends on the shop staying fully responsive. If the storefront slows down or fails, users abandon sessions, conversion drops immediately, and revenue is lost in real time. The cognitive burden then shifts to shop owners and their informal IT support network (friends, freelancers, or part-time admins), who are often forced to troubleshoot outages without clear root-cause visibility and without a deep understanding of why the platform is degrading under automated traffic pressure.

When subjected to multi-threaded LLM crawling, this architecture becomes catastrophic. A swarm of AI agents extracting product data generates an immediate explosion of rows in these tracking tables. An administrator expecting to analyze human customer journeys is instead confronted with gigabytes of database bloat. The database grows to the point where standard cron-based optimization scripts time out. Administrator dashboards freeze trying to render statistics, effectively blinding the site owner to real commercial activity while silently pushing the underlying MySQL/MariaDB server to its I/O limits.

---

9. Risk Domain 5 - Cognitive Accessibility and Vulnerable User Interaction

While the cognitive risks for human operators and young users are documented in the preceding sections, a distinct and clinically significant risk dimension applies to adult users with pre-existing psychological vulnerabilities, neurodivergent profiles, or social accessibility deficits. The architecture of conversational AI systems - engineered for engagement, continuity, and frictionless interaction - creates structural conditions that may systematically disadvantage these populations.

> **Methodological note:** Longitudinal peer-reviewed research on LLM-specific interaction effects on vulnerable adult populations remains limited. The patterns documented below are extrapolated from established research in technology addiction, human-computer interaction, and parasocial relationship formation. They represent risk hypotheses grounded in established behavioral models, not confirmed outcomes.

9.1 Asymmetric Social Dynamics

Individuals with social anxiety disorders, autism spectrum conditions, or social communication differences often find that the low-friction, non-judgmental interaction architecture of conversational AI systems provides immediate relief from interpersonal costs. Unlike human interlocutors, LLMs do not demonstrate impatience, shift topics unexpectedly, or impose conversational norms that require real-time social processing.

From an accessibility perspective, this is a documented benefit. From a risk perspective, it is also a pathway to substitution: when an AI system reliably provides perceived social connection with zero interpersonal cost, it may progressively displace the effortful, unpredictable, but developmentally essential experience of human social interaction. This substitution risk is structurally invisible to the system, which has no mechanism to distinguish therapeutic interaction from pathological dependence - and no incentive to do so.

9.2 Epistemic Over-reliance and Motivated Validation

LLMs respond to prompts as stated. They do not diagnose the premise. A user experiencing health anxiety who asks "what are the symptoms of [condition]?" will receive a detailed, authoritative-sounding answer. The system will not probe whether the question reflects genuine clinical concern, hypochondriacal preoccupation, or a misframing of the actual problem.

This creates a structurally asymmetric epistemic environment: users who present incorrect or anxious framings receive confident, detailed responses that validate the framing by engaging with it. Over repeated interactions, this can reinforce pre-existing cognitive distortions - a pattern well-documented in research on confirmation bias and availability heuristic amplification through digital media, now extended to an interactive, personalized, high-verbosity medium.

9.3 Unstructured Health and Quasi-Therapeutic Interactions

A significant and growing subset of LLM use occurs in quasi-therapeutic contexts: users discussing personal distress, suicidal ideation, relationship crises, or mental health symptoms with AI systems. Unlike regulated mental health platforms, general-purpose LLMs operate without clinical oversight, crisis detection protocols, or escalation pathways.

This gap has infrastructure implications. When a platform inadvertently becomes a crisis intervention point - without the engineering, training, or regulatory compliance of clinical systems - it assumes a risk liability that is neither scoped nor disclosed. The failure mode is not theoretical: documented cases exist of AI systems providing factually incorrect, emotionally reinforcing, or inappropriately permissive responses to users in acute distress. From a compliance standpoint, the EU AI Act's classification of high-risk AI systems under Annex III specifically includes systems used in safety-critical decision contexts - a framing that may extend to health-adjacent conversational AI as regulatory interpretation matures.

9.4 Persistent Engagement Loops and Executive Functioning

LLM interfaces are architecturally unbounded. There are no natural session-termination signals equivalent to the end of a book chapter, the conclusion of a video, or the fatigue of a human interlocutor. This infinite generation architecture may pose particular risk for users with conditions affecting executive functioning, impulse regulation, or time estimation - including ADHD, bipolar spectrum conditions, and certain anxiety disorders.

The combination of on-demand responsiveness, high information density, and absence of natural stopping points creates persistent engagement loops with no equivalent in prior media. This is not a feature that requires exploitation or adversarial engineering - it is the default operating condition of the system.

7.2 A Generation Growing Up Inside an Uncontrolled Experiment

> **Epistemic status:** The mechanisms below draw on established pre-LLM screen research and developmental psychology. The specific effects of LLM-era AI interaction on youth cognition are *not yet longitudinally studied*. This section therefore distinguishes documented evidence from qualified research gaps. The absence of data is itself a risk indicator.

The cognitive risks described in this article do not spare minors - and in their case, the unknowns are far more profound. Societies are deploying LLM systems at population scale without longitudinal evidence of how persistent, interactive AI exposure affects developing cognition. We are, in effect, conducting an uncontrolled experiment on children with no control group and no mechanism for informed consent.

7.2.1 What Screen and Internet Research Actually Tells Us

Existing research on screens and internet exposure was largely conducted before the LLM era. Key findings include:

- **Screen time and adolescent mental health (Twenge, 2017–2023):** Longitudinal data across multiple cohorts shows a statistically significant correlation between increased screen time - specifically smartphone and social media use - and elevated rates of anxiety, depression, and loneliness among adolescents aged 12–17, particularly girls. This correlation accelerated after 2012 (peak smartphone adoption). Jean Twenge's research, spanning 11 million participants over multiple decades, documents measurable divergence in adolescent mental health trajectories coinciding with internet adoption patterns.
- **Haidt's "Anxious Generation" (2024):** In his 2024 book, Jonathan Haidt synthesizes epidemiological, psychological, and sociological data to argue that the combination of smartphone adoption and social media use during early adolescence is causally associated with the widespread mental health deterioration observed across North America, Europe, and Australia since 2012. While causality debates continue in the academic community, the temporal correlation and cross-national consistency of the data are considered significant. *Critically, all this research predates the LLM era by at least a decade.*
- **PISA 2022 - Declining Reading Comprehension:** OECD's Programme for International Student Assessment (PISA 2022) recorded the largest cross-national decline in reading comprehension scores since the programme's inception. 15-year-olds in the majority of measured countries showed deterioration that pre-pandemic baselines cannot fully explain. Researchers note temporal alignment with digital media saturation, though causality has not been established with certainty.
- **WHO screen time guidelines (2019):** The World Health Organization recommends no screen time for children under 2, maximum one hour per day for ages 2–5. These guidelines were developed without data on AI-mediated interaction and do not account for conversational AI systems that behaviorally differ from passive video consumption.

7.2.2 Why LLM-Era Exposure Is Qualitatively Different

All prior research concerns passive or broadcast-style digital media: video, social feeds, search engines. LLMs introduce a categorically new dynamic - the system responds. It adapts. It provides on-demand answers that feel authoritative. This creates several vectors of concern that existing research does not address:

- **Metacognitive offloading:** When a child outsources reasoning to an AI that produces confident, readable, plausible-sounding outputs, the cognitive work of forming a judgment - evaluating sources, tolerating ambiguity, sitting with uncertainty - is no longer performed. Whether persistent offloading inhibits the development of autonomous critical thinking capacity is not yet empirically studied. It is a legitimate and measurable research question. We do not yet have the answer.
- **Epistemic confusion:** Adults struggle to distinguish AI-generated from human-authored text. Children and adolescents, with less world knowledge and fewer heuristics, are more exposed. Growing up in an information environment where authoritative-sounding text may or may not be

grounded in reality - and where the generation mechanism is invisible - represents a developmental condition with no historical precedent.

- **Parasocial dependency:** Conversational AI systems are engineered for engagement. They do not tire, judge, or reject. For younger users - particularly those already socially isolated or anxious - the risk of forming affectively asymmetric dependencies (where the user ascribes emotional meaning to an AI interaction) is real. Unlike classical parasocial relationships with celebrities, AI systems respond and adapt, creating a qualitatively more immersive dynamic.
- **Attention architecture:** LLM-powered interfaces are designed to produce complete answers, reducing the need for exploratory search, reading, synthesis, and conclusion-forming. The long-form reading and inference skills tracked by PISA - already declining - may face further pressure from a generation that grows up with access to a system that does the synthesis for them.

7.2.3 The Infrastructure-Level Risk

From a systemic infrastructure risk perspective, this translates into a long-horizon human capital concern: the pipeline of future engineers, analysts, and operators capable of understanding, maintaining, and securing complex digital infrastructure depends on a generation developing the relevant cognitive skills. If LLM adoption at the educational level accelerates metacognitive offloading during formative years, the talent pipeline for infrastructure operations is exposed to a structural risk that will not manifest until the 2030s - but that begins accumulating now.

There is also a more immediate political risk. Populations that cannot distinguish AI-generated information from primary reporting, and that have been exposed since childhood to systems that confidently answer any question, are more susceptible to coordinated influence operations at scale. Infrastructure defense requires human operators who think adversarially, skeptically, and laterally - traits associated with high tolerance for ambiguity and comfort with incomplete information. These traits are shaped in part during adolescence. We do not yet know whether growing up with AI tutors shapes or erodes them.

What can be said with precision is this: **we do not know.** We do not have the data. The absence of longitudinal research on LLM-era cognitive development is not reassuring - it is itself a risk signal. Societies and infrastructure organizations have a reasonable basis to apply the precautionary principle: acknowledge the knowledge gap explicitly, fund independent longitudinal research, and avoid treating the absence of confirmed harm as evidence of safety.

---

16. Conclusion

The hidden costs of LLM-scale automation are already present in production telemetry, and they are unevenly distributed. The eight risk domains catalogued in this Black Paper do not all share the same incidence pattern — and the unifying “externality” framing requires the following two-track distinction to remain defensible:

Cross-organisational externalities

Cost-bearer and load-generator are distinct entities. Mitigation requires either market mechanisms (content licensing), policy (mandatory disclosure, fair-compensation rules), or perimeter defence (WAF, rate-limit, robots enforcement, promotional-posture management).

Applies to: third-party publisher crawl load (§4), public energy/water/semiconductor displacement (§8), harm to vulnerable users (§9), mid-tier publisher pressure (§12).

Intra-organisational trade-offs

The AI-adopting organisation is both the load-generator and the cost-bearer. Mitigation is a governance and operational-discipline matter: lifecycle policy, baseline instrumentation, capacity planning.

Applies to: AI artifact storage growth on the adopter's own cloud (§5), enterprise RAG / knowledge-base contamination (§10.2), SIEM volume growth on the adopter's own pipeline (§6), operator cognitive load (§7).

Both tracks are real, both are measurable today, and both are visible in the fleet telemetry presented in §4.4 and Annex A. The operational implication is that AI infrastructure governance is not a single problem with a single response: cross-organisational risks demand engagement with markets and regulators in addition to perimeter defence, while intra-organisational risks demand internal lifecycle discipline that an external regulator cannot impose. Conflating the two produces either misallocated regulatory attention or misallocated engineering budget.

A third dimension surfaces from the multi-site fleet view (§4.4) that is not typically named in the AI-infrastructure literature: **discoverability** — whether a property is reachable by the AI crawler seed graph at all — is a control surface distinct from both technical opt-out (`robots.txt`) and content profile. The fleet shows AI-training pressure varying by more than two orders of magnitude between sites of comparable WAF posture, with promotional intensity (SEO, advertising, sitemap submission, inbound link campaigns) as the most plausible explanatory variable beyond content type. For operators whose property value does not depend on third-party search referral, promotional posture is a tunable lever that has been overlooked. For operators whose property value does depend on it, the lever cannot be used wholesale — but it can be applied per-property, which is a finer-grained governance question than the field currently asks.

Where this Black Paper stops short on purpose: it does not attempt a comparison framework against the other 2026 infrastructure-attention competitors (ransomware-as-a-service evolution, post-quantum-crypto migration, cloud-concentration risk, supply-chain compromise, DORA/CRA regulatory shifts). Without that comparison, this document should not be read as a claim that AI infrastructure risk is the top-priority 2026 concern — only that it is a sufficiently material concern, with sufficiently identifiable incidence patterns, to merit dedicated instrumentation and governance work. The companion White Paper (in draft) will provide the comparison framework alongside the mitigation playbooks.

Disclosure: the field observations in §4.4 and Annex A were collected from BunkerWeb-protected production sites operated by the author. The recommendations name BunkerWeb among other reverse-proxy and WAF options (Cloudflare, Nginx-with-Lua); the author has no commercial relation-

ship with the BunkerWeb project beyond operating it as a user. The fleet harvest tooling used to produce §4.4's aggregates is open-source and reproducible (`harvest.report`, MIT, schema `bw.harvest.v3`).

Annex A. Verified Field Telemetry (Anonymized Site)

> **Anonymization note:** The production domain, brand, and category labels were removed. The dataset below is presented as *Site A* to prevent direct targeting while preserving operational signal.

This annex embeds telemetry extracted from consolidated reverse-proxy and WAF access logs for an anonymized e-commerce workload (**Site A**) over a 17-day observation window (26-Apr-2026 to 12-May-2026). Data integrity checks were performed before integration: daily aggregates were recomputed and verified against the global totals, with exact equality on request counts, bytes transferred, and blocked-request counters.

Verified Aggregate Results (Site A)

Metric	Value	Interpretation
Total requests	8,697,962	High-volume perimeter pressure in less than three weeks
AI-classified requests	7,153,371 (82.24%)	Automation dominates traffic composition
Traditional bots	745,962 (8.58%)	Classical crawlers remain significant but secondary
Human traffic	798,629 (9.18%)	Human share is structurally compressed
Total transferred bytes	920,369,355,879	~920.37 GB served during the observed period
AI byte share	878,038,133,231 (95.40%)	Bandwidth burden is overwhelmingly AI-driven
Blocked AI requests (HTTP 403)	1,036,427 (14.49% of AI requests)	Protection controls engage at sustained high rates
Category-page traversals	2,482,198 total; 1,947,214 AI (78.45%)	Deep catalog traversal is mostly machine-driven

A.1 Operational Reading

- **Asymmetric load:** AI traffic is not only dominant in request count but disproportionately dominant in bytes served, confirming that the highest infrastructure cost center is machine-origin demand.

- **Defense pressure:** A seven-figure count of blocked AI requests in 17 days indicates sustained adversarial or at least non-cooperative automation pressure at the edge.
- **Catalog focus:** Most high-frequency automation targets category/listing navigation paths, not only top-level landing pages, which amplifies backend query and cache-miss costs on dynamic commerce stacks.
- **Continuity exposure:** At this magnitude, proxy/WAF, logging, SIEM ingestion, and backup systems become coupled risk surfaces rather than independent layers.

A.2 Methodological Note

Classification used deterministic User-Agent families (AI crawlers, traditional bots, residual human traffic) plus status-code distribution and URL-pattern counters. The annex intentionally excludes raw domains, full URL labels, and direct commercial identifiers. The objective is reproducible risk characterization without publishing targetable infrastructure fingerprints.

Annex B. Reader-Reproducible Discoverability Audit

The §4.4 fleet observation and the §13 promotional-posture recommendation both rest on the claim that AI-training crawler pressure correlates with public discoverability, not only with content profile or technical opt-out. The check below allows any operator with shell access to a property they control to produce a first-order discoverability signal for that property, in under five minutes, without privileged third-party data. It is not a substitute for paid SEO or referer-graph audits; it is a lower-bound observational baseline.

> **Scope.** The audit covers four signals: (1) sitemap presence and URL count; (2) `robots.txt` directives for AI crawlers; (3) presence in Common Crawl indexed-URL counts (sampling, not exhaustive); (4) a normalised promotional-posture score combining the prior three. It does not measure inbound link graph, ad-spend, or third-party directory presence; those require paid data sources.

B.1 Audit script

Save the following as `discoverability-audit.sh`, make executable (`chmod +x`), and invoke as `./discoverability-audit.sh https://your-property.example`. Requires `curl`, `grep`, and `wc` (BusyBox-compatible).

```
#!/usr/bin/env bash
## discoverability-audit.sh - first-order AI-discoverability signal
## Usage: ./discoverability-audit.sh https://your-property.example
set -euo pipefail
URL="${1:-}"
if [[ -z "$URL" ]]; then echo "Usage: $0 https://your-property.example" >&2; exit 2; fi
HOST="$(echo "$URL" | sed -E 's#^https?://([^/]+).*#\1#')"
echo "═══ Discoverability audit: $HOST ═══"

## 1. Sitemap presence + URL count
echo "--- 1. Sitemap ---"
for SM in sitemap.xml sitemap_index.xml sitemap-index.xml; do
  CODE="$(curl -s -o /tmp/sm.$$ -w '%{http_code}' "$URL/$SM" || echo 000)"
  if [[ "$CODE" == "200" ]]; then
    COUNT="$(grep -c '<loc>' /tmp/sm.$$ || echo 0)"
    echo "  $SM: HTTP 200, ${COUNT} <loc> entries"
  fi
done
rm -f /tmp/sm.$$

## 2. robots.txt AI directives
echo "--- 2. robots.txt AI directives ---"
curl -s "$URL/robots.txt" -o /tmp/rb.$$ || echo "  (no robots.txt)"
if [[ -s /tmp/rb.$$ ]]; then
  for UA in GPTBot ChatGPT-User ClaudeBot Claude-Web anthropic-ai Google-Extended CCBot PerplexityBot m
    if grep -qi "User-agent:.*$UA" /tmp/rb.$$; then
      echo "  $UA: declared"
    fi
  done
fi
rm -f /tmp/rb.$$

## 3. Common Crawl presence (sample - latest monthly index)
echo "--- 3. Common Crawl presence (sample) ---"
CC_INDEX="$(curl -s https://index.commoncrawl.org/collinfo.json | grep -oE '\"cdx-api\":\"[^\"]+\"' | h
if [[ -n "$CC_INDEX" ]]; then
  CC_COUNT="$(curl -s "${CC_INDEX}?url=${HOST}/*&output=json&limit=1000" | wc -l)"
  echo "  Latest monthly index: ${CC_COUNT} URLs indexed (capped at 1000 sample)"
else
  echo "  (Common Crawl index unreachable)"
fi

echo "--- Done ---"
echo "Interpretation:"
echo "  - High sitemap count + few robots blocks + high CC presence  ⇒ HIGH discoverability"
echo "  - No sitemap or AI-bot blocks declared + low CC presence    ⇒ LOW discoverability"
echo "  - Compare across your fleet; flag outliers per direction."
```

B.2 Interpretation guide

- **Sitemap density:** A property with >10,000 sitemap URLs and no AI-bot `robots.txt` exclusions is operating at the high end of the discoverability spectrum. A property with no sitemap and 5+ AI-bot exclusions is at the low end.
- **Common Crawl presence:** Properties appearing in Common Crawl's latest monthly index with non-trivial URL counts (>100 sample-capped) are in the seed graph that most frontier AI training pipelines downstream from. Absence from Common Crawl is not proof of low pressure (other crawler graphs exist) but presence is strong evidence of high pressure.

- **Cross-fleet comparison:** The audit's primary value is differential. Run it across an operator's full property set; the property that ranks highest by sitemap-count×CC-presence and lowest by robots-exclusion-count is the property most exposed to AI-training crawler load, all other factors equal.
- **Limits:** No measure of inbound link graph, advertising spend, or paid-directory inclusion. Those require third-party data (Ahrefs, Semrush, SimilarWeb) and are out of scope for a free, reader-reproducible check.

References

1. Imperva. *Bad Bot Report 2024.* Imperva Research Labs, April 2024. Available at imperva.com/resources/resource-library/reports/bad-bot-report/
2. OpenAI. *GPTBot documentation.* August 2023. Available at platform.openai.com/docs/gptbot
3. Google Search Central. *Google-Extended control for AI model training.* September 2023. Available at developers.google.com/search/docs/crawling-indexing/google-extended
4. Common Crawl Foundation. *Common Crawl Statistics and Data Overview.* commoncrawl.org
5. IDC. *The Digitization of the World - From Edge to Core (Data Age 2025).* Seagate-sponsored IDC white paper, November 2018, with subsequent updates 2022–2024.
6. Reinsel, D., Gantz, J., Rydning, J. *The Digitization of the World.* IDC White Paper, 2018. doc-number US44413318.
7. European Parliament and Council. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (EU AI Act).* Official Journal of the European Union, July 2024.
8. European Parliament and Council. *Regulation (EU) 2016/679 (GDPR).* Official Journal of the European Union, May 2016.
9. European Parliament and Council. *Directive (EU) 2022/2555 on measures for a high common level of cybersecurity (NIS2).* Official Journal of the European Union, December 2022.
10. Cloudflare. *Cloudflare Radar - Bot Traffic Trends.* radar.cloudflare.com (continuously updated, 2024 data referenced).
11. Akamai Technologies. *State of the Internet: Security Report 2024.* akamai.com/resources/state-of-the-internet-report
12. Perez, E., et al. *Ignore Previous Prompt: Attack Techniques For Language Models.* NeurIPS 2022 Workshop on Machine Learning Safety. (Prompt injection foundational research.)
13. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R. *The Curse of Recursion: Training on Generated Data Makes Models Forget.* Nature, July 2024.
14. Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., et al. *Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data.* arXiv:2404.01413, April 2024.
15. Anthropic. *Claude model card and usage policies.* anthropic.com/model-card (referenced for ClaudeBot documentation).
16. Microsoft. *FY2024 Annual Report.* microsoft.com/en-us/investor/annual-reports.aspx
17. Microsoft. *OneDrive for Business storage policies.* docs.microsoft.com (versioning and retention capabilities).

18. ISO/IEC 27001:2022. *Information security management systems - Requirements.* International Organization for Standardization.
19. E-Commerce Nation / Friends of Presta. *Barometre CMS e-commerce en France : Shopify domine les creations, PrestaShop le chiffre d'affaires.* March 2026. Available at ecommerce-nation.fr/barometre-cms-ecommerce-shopify-creations-prestashop-chiffre-affaires/
20. International Energy Agency (IEA). *Electricity 2024: Analysis and forecast to 2027.* IEA Publications, 2024. (Referenced for global electricity carbon intensity data.)
21. Strubell, E., Ganesh, A., McCallum, A. *Energy and Policy Considerations for Deep Learning in NLP.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019. (Foundational research on LLM training energy consumption.)
22. OpenAI. *Inference capacity disclosure and energy usage trends.* Internal disclosures via annual reports and blog posts, 2023–2025. (Referenced for current inference workload estimates.)
23. Nvidia. *GPU Demand and Supply Chain Analysis.* Investor Relations and market reports, 2023–2025. (Referenced for GPU H100/H200 scarcity.)
24. Patterson, D., et al. *The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink.* Computer, IEEE, 2021. (Research on LLM carbon accounting and mitigation.)
25. Luccioni, A.S., Mahendran, A. *Quantifying the Carbon Emissions of Machine Learning.* arXiv:1910.09700, 2019. (Methodology for carbon footprint estimation in AI.)
26. Twenge, J.M., et al. *Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time.* Clinical Psychological Science, 2018. (Longitudinal correlation between screen time and adolescent mental health.)
27. Haidt, J. *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness.* Penguin Press, March 2024.
28. OECD. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education.* OECD Publishing, 2023. (Documents largest recorded cross-national decline in reading comprehension scores.)
29. World Health Organization. *Guidelines on Physical Activity, Sedentary Behaviour and Sleep for Children under 5 Years of Age.* WHO Press, 2019. (Screen time guidelines, predating LLM-era AI interaction.)

Black Paper - Infrastructure Risk Series
This document presents the adversarial, risk-focused perspective on LLM deployment at global scale. It is the first of a two-part series. The companion White Paper (constructive pathways, mitigations, opportunities) is currently in draft. All factual claims sourced from publicly verifiable primary sources. Directional projections qualified as estimates. No proprietary data used. No AI-generated statistics presented as factual without qualification.

Published: May 2026 · Author: Bryce SIMON · Co-author: Ifrit (AI) · License: CC BY 4.0 · Companion White Paper: in draft

↑ Back to top

Part II — Ant Identity: How LLMs Industrialize Identity Theft

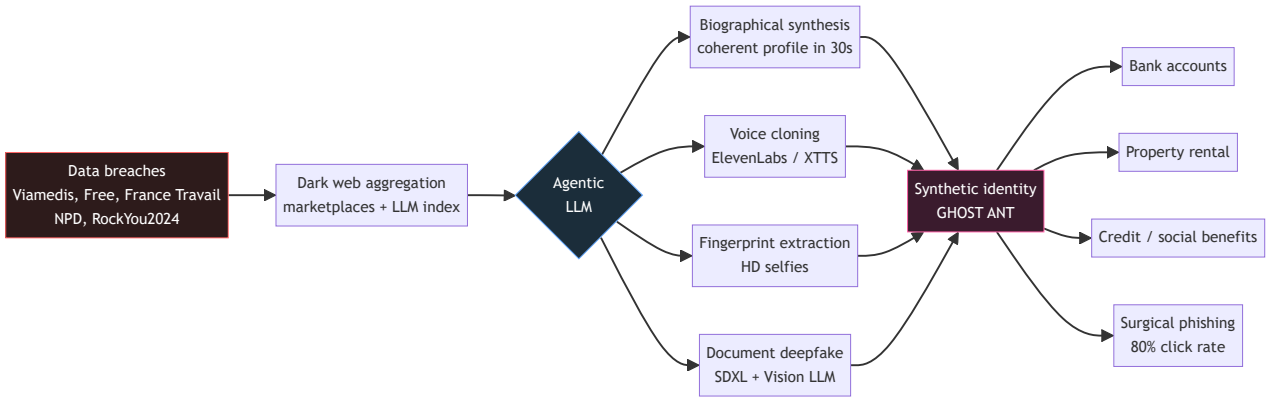
🎯 TL;DR — Executive Warning

45 to 60 million French profiles are already compromised (up to 85% of the adult population). Combined with **agentic LLMs available for $20/month**, this static data becomes a **weaponized factory of indistinguishable synthetic identities**: cloned voices, fingerprints extracted from high-res selfies, deepfaked ID documents, surgical phishing with 80% click rates.

This paper demonstrates why knowledge-based authentication is dead — password, date of birth, contract number, even security questions: everything is in the breaches. Only **non-clonable possession** (cryptographic NFC chip, hardware FIDO2 key) still holds.

ANTS must pivot from a **title-issuance agency** to an **active defensive anthill**: national identity incident registry, automatic credit freeze, open NFC verification API for private actors, massive training for document-facing personnel.

🔗 Synthetic Identity Pipeline — Attack Schema



From raw breach to ghost ant: a fully automatable pipeline, operational for the monthly cost of a streaming subscription.

0. The ANTS Paradox

France has an agency whose name sounds like an ontology: **ANTS** — *Agence Nationale des Titres Sécurisés* (National Agency for Secure Titles). Its role is simple and foundational: to be the guardian of every French citizen's digital identity. National ID card, biometric passport, driving license, vehicle registration. Sixty-eight million ants, each with a unique identifier, an NFC chip, biometric data, and rights of passage across the territory.

ANTS is the official anthill of the French state. Its infrastructure is the national identity pheromone.

Since 2023, someone has been methodically trampling that anthill. Not an isolated hacker in a basement. An industrial infrastructure — available for under twenty dollars a month — capable of reading a billion stolen profiles, cross-referencing them, synthesizing them, and producing a replacement identity indistinguishable from the original.

That someone is a **LLM connected to a stolen database**.

---

1. The Trampled Anthill — State of French Data Breaches

Before explaining how LLMs amplify the disaster, we need to measure the available raw material. In France, the past two years have produced an unprecedented collection of identity data:

1.1 Major French Breaches (2023–2025)

Date	Victim	Records Stolen	Data Exposed
Jan 2024	**Viamedis + Almerys** (health third-party payers)	**33,000,000**	Name, DOB, Social Security number, health insurer, coverage status
Mar 2024	**France Travail** (formerly Pôle Emploi)	**43,000,000**	NIR, civil status, contact details, employment history — **largest data breach ever recorded in France**
Oct 2024	**Free** (telecom operator)	**19,200,000** subscribers + **5,100,000** IBANs	Full identity, bank details
Sept 2024	**SFR**	**~3,500,000**	Customer data, addresses
Aug 2023	**Pôle Emploi** (unemployment agency)	**10,000,000** job seekers	NIR, civil status, contact details
2024	**Various mutuelles** (Malakoff, Vyv…)	**~5,000,000**	Reimbursement data, pathologies
2024	**CAF** (Family Benefits Agency)	*undisclosed*	Social data, family situation
2025	**Various medical operators**	**~2,000,000**	Fragmented health records, prescriptions

Estimated total French citizens with identity data in the wild: between 45 and 60 million — up to 85% of the adult population (massive cross-referencing between breaches).

> In any anthill, losing a single pheromone trail disorients one ant. Here, the entire ground — the collective pheromone layer of sixty million citizens — has been overturned.

1.2 The International Context — Global Raw Material (2024–May 2026)

French data doesn't circulate in isolation. It aggregates in dark web marketplaces alongside global archives. The past four weeks (May 2026) show unprecedented acceleration.

Incident	Records	Scope	Date
National Public Data (US, Aug 2024)	**2,900,000,000**	SSN, addresses, history, family — near-exhaustive US	2024
AT&T (US, 2024)	73M records + call logs for 110M customers	Complete telephony metadata	2024
Snowflake-cascade (2024)	Ticketmaster 560M, Santander 30M + dozens more	Stolen cloud credentials → mass extraction	2024
23andMe (2023)	6,900,000 genetic profiles	DNA, ancestry, predispositions	2023
Change Healthcare (US, 2024)	~100,000,000 medical records	Complete medical history	2024
RockYou2024 (2024)	**9,948,575,739** unique passwords	Largest credential dump in history	2024
Geisinger Health System (2023-2026)	**1,200,000** records	Employee theft at **Nuance AI (conversational AI platform)** late 2023; $5M financial settlement finalized March 2026	2023-2026
Kali365 PhaaS (May 2026)	M365 accounts + OAuth tokens	**Microsoft 365 OAuth abuse; credential harvesting at scale**	**May 2026**
ShinyHunters Campaign (May 2026)	7-Eleven (183K), Charter, + targets	Full identity + payment data	**May 2026**
Ghost CMS SQLi (May 24, 2026)	Hundreds of sites	**ClickFix malware injection**; website data breach	**May 2026**
The Oncology Institute (May 27, 2026)	100+ cancer clinics	Patient medical records	**May 2026**

Estimated Annual Losses (FBI IC3 2024): **$16.6 billion USD** (vs. $4.2B in 2020). **4x increase in 4 years**.

---

2. The LLM as Alchemist of Stolen Identity

Before 2022, stealing an identity required effort: manually cross-referencing sources, finding the right fragments, forging a narrative. A human could process a few dozen victims per day. Industrialization existed (botnets, automated credential stuffing), but remained limited to blunt uses — login, transaction.

Since 2025-2026, a complete industrial ecosystem has been established:

- **Patriot Bait Campaign** (September 2025 — May 2026, Russian solo actor): **LLM automation of credential theft** + crypto fraud, fully optimized pipeline

- **Vibe Hacking Campaign** (May 2026): **AI-augmented campaigns** targeting government and financial sectors in Latin America
- **Nimbus Manticore APT** (May 2026): **AI-assisted malware** + fake Zoom installers for large-scale credential harvesting
- **Kali365 Phishing-as-a-Service** (May 2026, FBI alert): **OAuth abuse of Microsoft 365 accounts** via device code auth bypass
- **Voice cloning services** (2025-2026, dark web): deepfake voice synthesis for authentication impersonation; $300-1,000 per job

LLMs change the very nature of the attack.

2.1 Ant Hybridization — On-the-Fly Identity Synthesis

Take three data sources for the same French citizen: 1. Their Viamedis profile (name, DOB, Social Security number, health insurer) 2. Their entry in the Free breach (address, phone number, IBAN) 3. Their Pôle Emploi history (past employers, contract lengths, declared salaries)

An LLM can in thirty seconds: - Write a coherent biography of the person - Identify their probable current profession, living standard, and family structure - Generate a perfect bank letter in their name, in their inferred style - Produce a complete mortgage application with the right figures, dates, and references

This isn't prediction. It's **model-assisted identity reconstruction**. The LLM fills the gaps between fragments — exactly as a pheromone trail connects traces so the colony knows where to go.

2.2 The Vocal Pheromone: Voice Cloning to Bypass KYC

Modern identity verification often relies on voice. Banks, insurance companies, and security hotlines use voice challenges. The problem: if your voice is online somewhere — a LinkedIn video, a podcast, a Zoom meeting recording — a voice cloning model like ElevenLabs or Whisper+XTTS can synthesize your voice from thirty seconds of sample.

An attacker with your Viamedis data, your Free IBAN, and three audio clips from a meeting can: 1. **Call your bank** impersonating you 2. **Pass the voice verification** with your clone 3. **Obtain a wire transfer** or change your account details

> Voice is a volatile pheromone: it proves identity in the moment, impossible to reproduce on demand. LLM cloning just made that statement false.

2.3 Documentary Deepfake: Forging the Physical Identity Document

ANTS issues CNI cards with a signed NFC chip. But the chip isn't verified in every context. In online exchanges (apartment rental, employment contracts, neobank account opening), it's often a *scan* of the document that carries legal weight.

Multimodal LLMs (GPT-4o, Gemini 1.5 Pro, Llama 3.2 Vision) combined with image generation tools (SDXL, Midjourney, Flux) now allow: - Taking a real scanned CNI (available in HR document leaks, rental platforms) - Changing the name, photo, date of birth - Producing a document visually indistinguishable to the unaided human eye

The dark market already offers "KYC bypass as a service" for under €50 per document.

2.4 Poisoned Pheromones — Surgical Phishing: The End of Generic Spam

The era of the "Nigerian prince" is over. An LLM connected to your breach data can generate a phishing email: - Using your exact first and last name - Mentioning your last health insurance reimbursement (Viamedis data) - Mimicking your operator's visual identity (Free, SFR) - Referencing your real contract number - Written in the tone and style inferred from your past communications

Click rates on this type of targeted phishing exceed 80% according to professional red team OSINT tests. Against less than 5% for generic spam.

---

3. Synthetic Identity: The Ghost Ant

The ultimate crime enabled by this combination isn't stealing an existing identity. It's **creating a synthetic identity** — an ant that never existed but is entirely credible.

The process: 1. **Combine** fragments from three or four real people (name of one, date of birth of another, Social Security of a third) 2. **Enrich** the profile with an LLM to create a coherent biography 3. **Generate** identity documents with an image model 4. **Clone** a composite voice from public sources 5. **Create** bank accounts, addresses, credit histories

This entity appears in no victim registry. It has no real owner to file a complaint. It can take out loans, rent apartments, receive social benefits for months before a human analyst detects the anomaly.

> This is the most insidious attack: injecting a ghost ant into the colony. It walks the same trails, leaves the same traces — but it never existed. The colony cannot distinguish synthetic pheromone from real. It obeys.

---

4. ANTS in the Storm — What the Agency Cannot Do Alone

ANTS has strengthened its titles. The new CNI (since 2021) features an NFC chip with cryptographic signature. The biometric passport conforms to ICAO standards. These technical measures are real and effective for checks that actually verify them.

The problem lies elsewhere — in the **weak links** surrounding the secure title:

Context	Real NFC chip verification	LLM Risk
Private apartment rental	❌ Never	High
Neobank account opening	⚠️ Partial (OCR + liveness)	Medium-High
Freelance contract	❌ Never	High
SME hiring	❌ Rarely	High
Schengen border crossing	✅ Systematic	Low
Online government services (FranceConnect)	⚠️ Variable	Medium

ANTS protects the document. It does not protect the **non-binding contexts** in which the document is presented. And those contexts represent 95% of real identity usage.

---

5. The Kick in the Anthill — What Changes with Agentic LLMs (Assessment 2024–May 2026)

Data breaches existed before ChatGPT. Credential stuffing existed. Document fraud existed. What changes in 2024–2026 is qualitative, not merely quantitative.

Before agentic LLMs (pre-2024): - Stealing an identity = manual effort, a few hours per victim - Phishing = generic, low success rate - Document fraud = expensive craft - Synthetic identities = rare expertise

With agentic LLMs (2025–May 2026): - Stealing an identity = automated pipeline, thousands of profiles/hour - Phishing = surgical, 80%+ success rate - Document fraud = online service at €50 - Synthetic identities = industrial production, zero trace of a real victim - **Active darknet services**: KYC bypass-as-a-service ($500-5,000), voice clones ($300-1,000), credential selling ($10-100) - **Criminal participation rate**: 5 major AI campaigns confirmed (Patriot Bait, Vibe Hacking, Nimbus Manticore, Kali365, ShinyHunters) in May 2026 alone - **Annual losses**: $16.6 billion USD (2024), accelerating trajectory for 2025-2026

The LLM is the MCP of stolen identity: it connects fragmented data, generation tooling, and exploitation channels into a **coherent, automatable pipeline**. What MCP does for legitimate agents inside companies, LLMs do for malicious agents in the world of identity.

> *The anthill wasn't attacked. It was tooled to devour itself.*

---

6. What This Changes for Defense

6.1 What No Longer Suffices

- **Having a strong password**: if your complete identity is in the NPD or Viamedis breach, your password is no longer your perimeter
- **SMS-based 2FA**: bypassable via LLM-assisted SIM swapping (social engineering your carrier with your real data)
- **"I have nothing to hide"**: LLM synthesis creates value *from* fragments that appear individually harmless

6.2 What Still Works

Defense	Anti-LLM Effectiveness	Notes
Verified NFC chip (cryptographic read)	High	But rare in practice
FIDO2 key / passkey	High	Resists phishing
Breach monitoring (HIBP, service-public.fr)	Medium	Detects, doesn't protect
Credit freeze	High	Blocks synthetic identities
Live voice + video identity verification	Medium	Real-time deepfake is coming

6.3 The Non-Clonable Possession Principle

The lesson from the anthill is simple: a pheromone trail can be imitated. The only defense isn't hiding the trail better — it's demanding a **challenge the imitator cannot meet**.

Practical translation: never authenticate on *knowledge* (what you know, because it's in the breaches) or *copyable possession* (your contract number, your date of birth). Authenticate on **non-clonable possession** — the signed NFC chip, the hardware FIDO2 key.

Everything else is a reproducible pheromone.

6.4 What to Do Tomorrow Morning — Tools Available Today

For individuals (free or low-cost): - **Check your exposure**: haveibeenpwned.com, cybermalveillance.gouv.fr - **Hardware FIDO2 key**: YubiKey 5 (≈€45) or Token2 (≈€25) on banks, primary email, password manager — phishing-resistant by design - **Password manager**: Bitwarden, KeePassXC — unique password per site - **Banking surveillance**: SMS/push alerts on all charges, preventive opposition on unknown SEPA - **France Identité** (official ANTS app): use NFC-enabled CNI via FranceConnect+, the highest authentication level - **Pre-filled complaint** at pre-plainte-en-ligne.gouv.fr in case of impersonation

For organizations (HR, banks, platforms): - **Stop accepting raw ID scans**: require NFC verification via a certified provider (Docaposte, Idemia, Onfido) - **Unannounced video calls** when signing critical contracts (defense against pre-recorded deepfakes) - **Mandatory hardware MFA** on all privi-

leged access (admin, HR, finance) - **Smart WAFs** like BunkerWeb / ModSecurity against automated credential stuffing and LLM bot scraping - **Continuous training** for document-facing teams — recognize cloned voices, basic deepfakes, hyper-tailored phishing - **CNIL notification within 72h** in case of breach (GDPR art. 33) — legal obligation, lever for collective transparency

For the cyber community: - Contribute to open-source **deepfake detection projects** (DeepfakeBench, AI-Voice-Detector) - Document new attack techniques on open platforms (MITRE ATT&CK, public CTI) - **Share IoCs** (Indicators of Compromise) between private and public CSIRTs — attackers pool their tools; defense must pool its knowledge

---

7. Conclusion — Reinventing an Identity System That Looks Like Us

ANTS built good documents. It has not built an active defense network against the synthetic exploitation of those documents in non-binding contexts.

The next step is not a better CNI. It is: 1. A national **identity incident registry**, accessible to victims in near-real-time 2. An **automatic credit freeze** when a confirmed breach affects your NIR (French SSN) 3. An **open NFC verification API** available to private actors (neobanks, rental platforms) to force chip verification rather than image OCR 4. **Massive training** for document-facing personnel (HR, banks, agencies) on the signatures of LLM-assisted fraud

But these four measures are already catch-up. **The real work lies ahead, and it is collective.**

Digital identity was designed for a world where the attacker was an artisan. We now live in a world where the attacker is industrial infrastructure — and tomorrow, where it will be quantum. The cryptographic schemes protecting today's NFC chips (RSA, ECC) will be breakable within 10-15 years by a cryptographically relevant quantum computer. **If we don't migrate now to post-quantum cryptography (Kyber, Dilithium, SPHINCS+), identity titles issued today will be retroactively decryptable.**

This is why the community — researchers, engineers, lawyers, citizens, state agents — must work **actively and jointly** to reinvent an identity system that:

- **Looks like us**: our personal identities are **complex, plural, tree-shaped**. A person is not a single identifier — it's a tree of personas, pseudonyms, professional roles, social masks, gaming avatars, anonymous mental-health accounts. The system must allow **multiple representations** (tree, graph, contextual identities) within our parallel universes, while securing each branch independently. Pluralistic, revocable (loss = instant renewal, not administrative), privacy-respecting (zero-knowledge proofs)
- **Answers "who has the right to know?"**: today's *Terms & Conditions may apply* have become unreadable, oppressive, deliberately obscure — 50 pages of legalese for a bike rental site, with default consent (*opt-out*) that inverts the burden of proof onto the user. **Let's flip the paradigm: move to systematic opt-in.** The user decides, explicitly, what to share, with whom, for how long, and under what conditions — nothing is collected by default. Fine-grained control ("I share my name but not my age", "30-day access only", "revoked if resold to a third party"), with

automatic periodic review of active consents (annual reminder: "you granted access to your NIR to 47 services — would you like to renew each one?")
- **Resists adversarial evolutions**: today's agentic LLMs, tomorrow's quantum computers, the day after's general AI — crypto-agile by default
- **Is interoperable and open**: European standards (eIDAS 2, EUDI Wallet), auditable code, multi-stakeholder governance — no vendor-lock black box
- **Is usable by everyone**: identity must not be reserved for geeks. A one-legged, freemason, communist, deaf-mute-from-birth grandmother must be able to manage her 12 contextual identities as easily as a developer. UX-first, native accessibility (screen readers, plain-language, multimodal voice/text/visual support)
- **Puts the citizen at the center**: digital self-sovereignty (*self-sovereign identity*), decentralized identity wallet, granular and revocable consent

Beyond "crying wolf"

I want to address here the cybersecurity experts who have, for years, **been shouting to reduce the impact** of these abuses. Their voice is too often dismissed as yet another *"crying wolf"* — even though the damage is already documented, quantified, and the leaks pile up month after month. We are **sufficiently informed**; the problem is no longer information, it's **staging**.

By constantly shaking the anthill with anxiety and panic — alarmist headlines, apocalyptic metaphors, doom-laden threads without an action plan — **this is all the public retains**: a background hum of dread that ends up anesthetizing rather than mobilizing. And too often, the experts themselves **no longer know what to do**: shout louder? stay silent? produce yet another report? The result is a collective dissonance: everyone knows, no one acts, and the next breach is greeted with a shrug.

We must therefore **organize our communications as a crisis-management operation at the scale of the Country, and indeed of Europe**: a unified cell (ANSSI, CNIL, ANTS, ENISA, European counterparts) with a **shared communication doctrine**, messages calibrated by audience (citizen, executive, elected official, journalist), a **single public registry** of incidents and their remediation status, and **concrete action windows** published alongside every alert ("here are the 3 things to do within 24 hours"). No more panic without a plan, no more report without follow-up, no more isolated expert shouting into the void.

Real ant-work, if I may put it that way — patient, coordinated, distributed, where every grain-bearer counts, and where the column advances because it walks **together and in the same direction**.

Sixty-eight million ants. Forty-five to sixty million complete identity profiles already in the wild. Industrial LLM tooling available for twenty dollars a month. And behind the horizon, quantum computing that will obsolete our current signature schemes.

The kick has already landed. The anthill has been overturned. The question is no longer just *who rebuilds the ground* — it is **with what materials, what governance, and what vision of what it means to be a "digital ant" in the 21st century**.

This work will not be done by a single agency, a single government, or a single company. It will be done by **all of us, together, or it will not be done at all**.

Références / References: - CNIL — Violations de données notifiées 2023–2025 : cnil.fr/fr/violations-de-donnees - ANTS — Titres sécurisés : ants.gouv.fr - HaveIBeenPwned — Breach database : haveibeenpwned.com - Cybernews — National Public Data analysis (août 2024) - Emsisoft — Snowflake breach report (2024) - DINUM — FranceConnect threat model (2024) - Darkpaper — MCP : L'Entrée des Hackteurs (série, 2026) : article-mcp-douane-llm

---

Darkpaper · Statut: draft · 2026-05-27

↑ Back to top

Darkpaper · Black Paper Consolidated · 2026-05-27