The 4,217 Logins Nobody Approved

Measuring shadow AI from the outside in

CTRL ALT Elite · Fortune 100-caliber infrastructure, operations, and security leadership

By René López, Founder and Principal · ORCID https://orcid.org/0009-0008-4993-0088

The Number Your Board Has Never Seen

Every month, employees at your company log into consumer AI tools thousands of times. Not the enterprise tools you licensed, governed, and negotiated data terms for. The free and personal tiers: the ones where the terms of service say, in plain language, that your inputs may be used to train the model.

Most executives have never seen this number for their own organization. Not because it is hard to measure, but because nobody is looking. The traffic is encrypted, the tools live in a browser tab, and the people using them are not trying to hide. They are trying to get their jobs done. Your analysts, your marketers, your junior engineers: the citizen AI developers who found a faster way and never thought to ask whether the customer list they pasted was leaving the building.

> *You had 4,217 logins to consumer AI tools from your corporate network last month. Here is what else we can tell you.*

That sentence, with your real number in it, is the most effective AI governance conversation starter we have found. This paper explains how the number is produced, why it is more honest than the alternatives, and what to do once you have it.

The Wrong Question: Did the Model Learn Our Data?

The first instinct after an incident is forensic: can we prove the model absorbed our information? The honest answer is that nobody can, reliably. Membership inference and extraction attacks produce probabilistic signals that degrade badly against large frontier models with deduplication and alignment layers. A vendor selling you a scan that claims to detect whether a specific prompt was trained into a specific model is selling confidence, not evidence.

The uncomfortable corollary is also the useful one: because you cannot prove your data was learned, and you cannot prove it was not, the only defensible posture is to assume that anything sent to a consumer tier is gone. That reframes the problem from forensics, which is unwinnable, to exposure measurement, which is entirely winnable.

The Right Question: What Is Leaving, From Whom, To Where?

Exposure is observable. It lives in three layers of telemetry, each with a different cost and fidelity:

1. **Identity:** which corporate accounts authenticated to which AI tools. Your identity provider already knows every SSO login. DNS logs reveal consumer logins that bypass SSO entirely. This is the cleanest signal and requires no packet inspection at all.
2. **Volume:** how much data moved. Encrypted traffic still discloses bytes transferred. Upload asymmetry is the tell: a session that pushes forty megabytes up to a consumer AI endpoint is a document dump, not a chat. Egress volume by destination is your leakage proxy.
3. **Content:** what was actually in the prompts. This requires TLS inspection or an endpoint agent, carries real privacy and legal weight, and should be a deliberate later step, not a starting point.

Most organizations need layer one and layer two to change behavior. Layer three is for regulated environments and confirmed problem areas. Critically, the first two layers can be stood up remotely, in days, with no hardware and no agents.

The Remote Architecture: Why Cloudflare Changes the Math

Historically, this visibility required an on-premises proxy or a CASB deployment measured in quarters. That barrier is gone. Cloudflare Zero Trust, specifically Gateway, provides the entire measurement plane as a cloud service, and the onboarding step is a DNS change:

- **Point office DNS at Gateway.** Change resolver forwarders at each site to Cloudflare's Gateway endpoints. From that moment, every DNS query from your corporate ranges is logged, and Cloudflare's application catalog classifies AI destinations natively: it already knows which hostnames belong to which AI product, and distinguishes categories of AI services out of the box.
- **Cover remote workers with WARP.** The WARP client extends the same visibility to laptops off the corporate network, which is where CIDR-based monitoring traditionally goes blind. Deployment is an MDM push.
- **Export with Logpush.** Gateway logs stream to R2 or S3 automatically. A lightweight reporting job, running as a Cloudflare Worker or a simple scheduled script, aggregates logins, sessions, and destinations into the monthly scorecard. No SIEM required to start.
- **Escalate only when justified.** If the snapshot reveals a hot spot, Gateway's HTTP policies with TLS inspection, and its CASB integrations, are already in place. You turn a dial rather than run a procurement cycle.

The result is a shadow AI assessment that is genuinely remote: no appliance shipped, no server racked, no agent required for the first phase. A client can authorize it on a Monday and see their number the following Monday.

The Scorecard: Numbers That Move a Board

Raw logs persuade nobody. The deliverable is a one-page monthly scorecard built on four measures:

- **Consumer AI logins from corporate identity or network,** total and by department. This is the headline number.
- **Estimated data egress to consumer tiers,** in bytes uploaded, trended month over month. Volume is the leakage proxy that duration is not: ten minutes asking a chatbot about spreadsheet formulas is harmless, ninety seconds pasting a customer file is not.
- **Consumer versus sanctioned ratio:** of all AI activity, what share flows to tools with enterprise data terms versus tools that train on inputs. This is the single best measure of whether your governance program is working.
- **Top unsanctioned destinations,** so the sanctioned catalog can grow toward what people actually want, instead of what a committee guessed.

> **DELIBERATELY ABSENT: TIME SPENT**
>
> *Duration metrics read as productivity surveillance, poison the relationship with the workforce, and prove nothing about leakage. Measure the data, not the person.*

Counting Tokens Without Seeing Them

Clients reasonably ask how token consumption can be measured when the traffic is encrypted and the providers report nothing to the network. The answer is a ladder of four methods, each with a different fidelity, and each mapped to a different level of client authorization:

1. **Estimation from egress volume.** With DNS and flow telemetry alone, bytes uploaded per session are visible even though content is not. Discounting roughly a third for protocol overhead and applying the standard four characters per token, a 100 kilobyte upload represents an estimated 15,000 to 18,000 tokens of prompt

material. Labeled honestly as an estimate, this is sufficient for the scorecard, because the value is in the trend, not the fourth digit.

2. **Exact counts through TLS inspection.** Where Gateway HTTP inspection is authorized, the actual prompt payloads are visible and can be tokenized precisely, for both prompts and responses. This is a deliberate escalation with legal review, not a default, because it means reading content.
3. **Native metering through Cloudflare AI Gateway.** For API traffic, the citizen developer writing scripts against OpenAI or Anthropic, calls routed through AI Gateway are logged with exact token counts, model, and cost, because providers return usage data in every API response. Pair it with a Gateway policy that blocks direct API endpoints, and ungoverned API usage becomes governed usage with perfect metering, by design rather than by detection.
4. **Provider administrative reporting.** Enterprise tiers of the major AI tools expose seat activity and consumption through their admin consoles and APIs. This supplies the sanctioned side of the ledger with no inference at all.

Reconciled together, these four sources produce a single consumption picture: how many tokens flowed to consumer tiers that train on inputs, versus governed tools that do not, org-wide and by department.

> *Tokens measure volume, not sensitivity.*

One caution belongs in every readout: a million tokens of public documentation pasted into a consumer tool is less dangerous than two hundred tokens of customer records. Token counts are the trend line; data classification is the risk weighting. A mature scorecard carries both.

From Measurement to Behavior: The Education Layer

The purpose of the number is not punishment. Citizen AI developers are your most motivated builders; the goal is to redirect them, not suppress them. Three interventions, in order of impact:

1. **Just-in-time redirection.** Gateway policies can present a branded page at the moment someone opens a consumer AI tool: this tier trains on your inputs, here is the sanctioned alternative, one click to get there. Intercept at the moment of use beats any training deck ever written.
2. **A sanctioned path that is actually good.** Shadow AI is a demand signal. If the approved tool is slower, gated, or worse, the scorecard will not improve. Provisioning enterprise tiers with proper data terms is half of every engagement.
3. **Scorecard transparency.** Publish department-level numbers internally. Nothing changes behavior in a leadership team faster than a peer comparison.

What Honest Looks Like

Our position, stated plainly: no one can scan a frontier model and tell you whether it learned your data. Anyone who says otherwise is guessing with a dashboard. What can be done, remotely and quickly, is to measure exactly how much of your information is flowing to services that reserve the right to learn from it, show you who is sending it, and give those people a better path.

> *Assume the model learned it. Then make sure there is less to learn next month.*

Work with CTRL ALT Elite

CTRL ALT Elite LLC provides fractional CIO, CTO, and CISO leadership to Series A and B startups and PE-backed operators. Founder and Principal René López brings more than 25 years of Fortune 100 infrastructure and operations leadership across travel, aviation, healthcare, and financial services. Shadow AI exposure assessments are delivered remotely. Start the conversation at ctrlaltelite.io.

CTRL ALT Elite
reboot@ctrlaltelite.io · 954-516-6909 · ctrlaltelite.io · Pompano Beach, FL
René López, Founder and Principal · ORCID https://orcid.org/0009-0008-4993-0088
Disrupt · Innovate · Elevate

© CTRL ALT Elite. Shareable in full with attribution. The 4,217 figure is illustrative. Vendor-neutral; Cloudflare is named as one example of the measurement plane, and no product is endorsed.