FORESIGHT FUTURE INSTITUTE · PRELIMINARY INVESTIGATION

Shadow Harvest

Agent Traces and the Infrastructure of Delegated Disclosure

> *One of the most context-rich records of your work can now leave the building before any breach — carried outward through a helpful agent, as a by-product of getting the job done.*

— ABSTRACT

Abstract.

Generative-AI agents have turned one of the most context-rich records an organization produces — the working trace of a task — into something that can become a transmitted operational record. This does not require a breach; it is delegated disclosure, and it happens by design.

This preliminary investigation maps the architecture of that disclosure and its consequence for data custody. Drawing on independent technical studies, commercial telemetry, first-party provider analyses and regulator findings, it establishes that the overall flow of context to generative-AI services is large and growing; that resellers of model access can be opaque and can misstate the model; that first-party studies indicate increasing delegation of operational decisions; and that an independent controlled benchmark found substantially higher leakage through internal coordination channels than through final outputs. It then locates the structural problem: an application-layer intermediary occupies a privileged position to correlate the transmitted trace with its routing, while the customer may be unable to establish which participant occupied that position — leaving custody unverifiable.

The report is deliberately bounded. It does not claim a proven market in stolen traces; it separates what is observed from what is inferred and what remains an open hypothesis, and it sets out a timestamped pre-analysis protocol to be publicly deposited before the first observation. The architecture itself — disclosure delegated, correlation concentrated, and the conditions for unverifiable custody established — is the finding.

> *The vendor tracks capability theft. The customer may have no traceable chain of custody.*

— CONTENTS

Contents.

Throughout, every load-bearing claim carries an evidence label — OBSERVED, ADJACENT, INFERENCE, OPEN — and a superscript reference to the Source Register (Appendix F).

— KEY TERMS

How to read this report.

Five terms carry the argument. They are defined here so the chapters can move quickly.

Delegated disclosure	The transmitted part of a task becoming available to a provider or intermediary as the output of a chain — goal, permission, selection, serialization, transmission, routing, retention — in which the user initiates the goal but does not author every subsequent context selection or disclosure.
Provider-visible trace	The transmitted operational record of an agent session: messages, excerpts, tool calls and results, outcome signals and metadata. Not the model's hidden reasoning, and not necessarily the full local trajectory.
Correlation junction	An application-layer intermediary positioned to line up the transmitted trace with the routing and fulfilment decisions made on its behalf.
Unverifiable custody	The condition in which a user cannot independently establish identity, path, transformation, purpose, retention/reuse or deletion/enforceability for transmitted data.
Evidence label	Each claim carries a tier and two attributes — evidence strength (A independent / B first-party / C single case / D secondary) and transfer distance (0 same population & mechanism → 3 analogy). A conclusion never inherits the class of its source.

> A note on reading the ledger: a strong, direct finding about one named service (strength A, distance 0) does not make "the same is true of opaque proxies" equally strong — that is a fresh inference at greater distance. The labels travel with the claim, not the citation.

— PROLOGUE

The Useful Agent.

An analyst opens a coding agent and types a single instruction: *“Reconcile last quarter’s figures against the contract terms and draft the board summary.”* What happens next is not one action but a cascade. The agent reads the spreadsheet, opens the signed agreement, greps the repository for the billing logic, runs the failing test, inspects the error, and assembles a clean answer. The analyst authored the goal. The agent executed much of what followed.

This is becoming an ordinary shape of agent-assisted knowledge work, and it is genuinely useful. It is also the quiet relocation of one of the most context-rich records an organization produces. To be helpful, the agent must see the material; to see it, the material must be selected, serialized into a model payload, and transmitted across whatever boundary separates the user’s machine from the system that fulfils the request. None of this requires a breach. Nobody needs to pick a lock. At the point of transmission, the data has not been stolen — it has been delegated outward, one helpful step at a time, by a worker who reasonably believed they were simply doing the job faster.

The conventional security imagination still pictures exfiltration as an intruder dragging data *out* across a guarded perimeter. That picture is now inverted. The high-value material moves outward at the user’s own initiative and, increasingly, at the agent’s own discretion — before any incident, without any incident, and disguised as productivity. This report is about what that transmitted record contains, who is positioned to see it, and why no single party in an opaque chain may be able to provide a complete, independently verifiable answer to a simpler question: once it has left, what becomes of it?

> *We do not claim a market in stolen work-traces has been proven. We claim something more durable: the transfer of high-value context now happens by design, the parties who could account for it are not the parties who can see it, and the user may have no verifiable custody chain by which to establish its fate.*

— THE ARGUMENT, IN ONE BREATH —

productive egress → provider-visible trace → delegated disclosure → correlation junction → unverifiable custody

— FRAMING

One infrastructure, two viewpoints.

In February 2026 a frontier lab disclosed that a single proxy network had run more than twenty thousand fraudulent accounts, deliberately blending industrial model-distillation traffic with unrelated customer requests to evade detection.[6] The disclosure was written, reasonably, from the vendor's vantage point: the harm it names is capability theft. But the same infrastructure can be read from the other end of the wire — from the position of a customer whose unrelated request was routed through the same network.

VIEWPOINT A · THE PROVIDER

What the vendor sees

› Capability theft — reasoning and outputs distilled into rival models

› Account abuse — bulk fraudulent registration, evasion of KYC

› Model extraction at industrial scale

› A threat to frontier economics and export-control objectives

VIEWPOINT B · THE CUSTOMER

What the customer faces

› Unknown routing — which model actually answered?

› Unknown retention — is the request stored, and for how long?

› Unknown enforceability — can any stated limit be verified?

› Correlation exposure — payload, account history, payment metadata and route may converge in one place

ADJACENT EVIDENCE The proxy network above is a vendor-described *adversarial* operation. It is an entry point into the custody question — not evidence that all third-party intermediaries operate this way. We use it to flip the camera, not to characterize the market.

> *The vendor tracks capability theft. The customer may have no traceable chain of custody.*

— SECTION 01

Productive data egress.

User-initiated data leaving the managed perimeter is not new. Shadow IT moved documents into webmail, personal cloud drives and online translators long before generative AI. What AI changed is not the existence of the flow but its character: disclosure became the precondition of a useful result rather than an incidental violation, and the potential informational density of a disclosed session increased.

The scale is real but must be read precisely. Through 2025–2026 the number of SaaS generative-AI users roughly tripled, and the volume of prompts sent rose about sixfold — from a median near three thousand to eighteen thousand per organization each month.[3] Over the same period the share of users reaching these tools through *personal* accounts fell from 78% to 47%, while organization-managed access rose from 25% to 62%. The two categories overlap — a single user may use both. So the headline is not “shadow AI keeps growing unchecked.” It is subtler: *the personal-account share is shrinking even as the absolute flow of context expands rapidly.*

PROMPT VOLUME	PERSONAL-ACCOUNT ACCESS	SAAS GENAI USERS
×6	**78 → 47%**	**×3**
~3,000 → ~18,000 / org / month	org-managed 25 → 62%	tripled over the period

Source. Netskope, *Cloud & Threat Report 2026.* OBSERVED — independent vendor measurement.

Three different things are being measured across this report, and the central error to avoid is fusing them into one number. The size of the flow is established for ordinary shadow AI — but that flow runs largely through *official* services (ChatGPT, Gemini, Copilot), not necessarily through opaque intermediaries. The opacity of resellers is established separately, for a different and smaller population. And the delegation of operational context is established for agentic systems, mostly from first-party

telemetry. Each leg is real; their intersection is not yet measured. The map below keeps them honest.

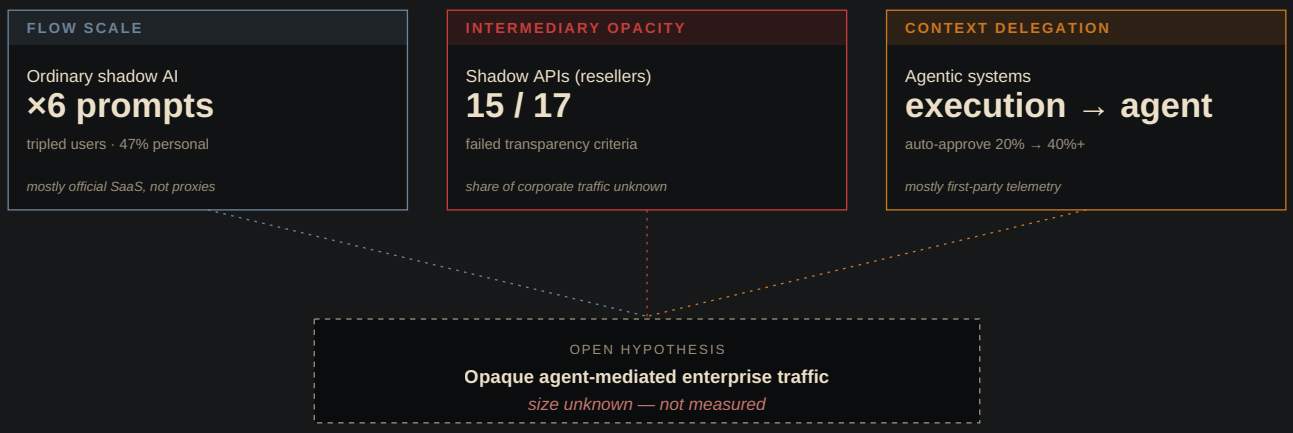


Figure 1. Three separately observed trends, each resting on its own population and method. The highlighted intersection — the share of delegated work-traces that passes through opaque intermediaries — is the quantity no current dataset establishes.

Naming that empty box is not a weakness of the investigation; it is one of its findings. We know the flow is enormous, we know intermediaries can be opaque, and we know agents now select operational context on their own. What no one can yet state is the size of their overlap — and a study that pretends otherwise would be measuring a market it has not seen.

— SECTION 01 · CONTINUED

The scale, precisely.

Before agents, user-initiated egress was already large — and already concentrated in the most sensitive material. Three separate observations establish the baseline.

SENSITIVE CONTENT TO PERSONAL ACCOUNTS

~51% / ~55%

source code · R&D — 98% via copy-paste[5]

AMONG BREACHED ORGANIZATIONS

13% · +$670K

AI-related breach; 97% lacked AI controls[4]

SHADOW-AI INVOLVEMENT

~1 in 5

of breaches in the cohort touched shadow AI[4]

Reading these. The percentages on the right are *within the breached cohort*, not the whole economy. OBSERVED for each measurement; ADJACENT when read as evidence about agent-mediated flow specifically.

The behaviour predates agents. In one 2024 measurement across roughly three million workers, about half of all source code and more than half of R&D material sent to generative-AI tools went to *personal* accounts, and almost all of it moved by copy-paste — outside any managed channel.[5] The pattern is not exotic misuse; it is the ordinary friction-reduction of people trying to work faster.

> CASE — BEFORE THE AGENT ERA
>
> Within twenty days of permitting an AI assistant internally, one large manufacturer recorded three separate disclosures — source code and internal meeting notes pasted into a public chatbot.[15] No external intrusion was required: just useful tools and sensitive context meeting at the keyboard. Agents do not introduce this behaviour; they can automate and extend it.

— SECTION 02

Anatomy of a provider-visible trace.

A chat prompt is a sentence. An agent session is a record. To work, an agent does not merely receive a question; it accumulates the working material around it — and the portion that crosses the boundary to be fulfilled can be substantially richer than a standalone prompt, even when much of the agent's activity stays local.

Precision matters here, because it is where a careless argument gets defeated. The intermediary does not necessarily see the model's hidden reasoning, and it does not always see the agent's full local trajectory: a file may be read on the device and only an excerpt sent; a shell command may run locally; secrets may be redacted by middleware. What it does see is the *transmitted* operational trace — and that alone typically contains:

COMPONENT TRANSMITTED	WHAT IT REVEALS
User messages & task framing	intent, goals, what "done" means
Attachments & transmitted excerpts	the actual proprietary content
Model-requested tool calls	intended actions (requested ≠ executed)
Tool results returned to context	systems actually reached; data, errors, structure
Tests, accepted changes & verified outcomes*	potential evidence of what succeeded — usable as a reward signal
Approvals & diffs	operational & contextual signals
Metadata (timing, account, sequence)	correlation handles

* when actually transmitted. INFERENCE Components vary by agent architecture; the table describes what is *commonly* present in a transmitted trace, not a guarantee for every system.

The right way to make the "richer unit" claim is not that the session wins on every axis — an ordinary chat may carry a large attachment, and not every agent has tool

access or an outcome label. The defensible claim is structural: a single agent trace can combine more of these dimensions in one record.

DIMENSION	ORDINARY PROMPT	PROVIDER-VISIBLE AGENT TRACE
Context volume	may carry a document or excerpt	may accumulate multiple artifacts and tool results
Sequentiality	often episodic	commonly multi-step, stateful
Tool / system access	optional	structurally common in agentic workflows
Outcome signals	uncommon	sometimes present (tests, validated task completion, accepted changes or explicit feedback)
Operational applicability	task-dependent	often closer to executable organizational action

Figure 2. A structural comparison, not a measurement. Disclosure is the precondition of usefulness; the session is sequential; and some trajectories carry a built-in outcome signal — together describing what a person knows, what they want, and how they act.

— SECTION 02 · CONTINUED

A session, dissected.

To make the transmitted trace concrete, here is a single, *illustrative* agent task — not a real capture — reduced to what crosses the boundary at each step.

#	AGENT STEP	WHAT CROSSES THE BOUNDARY
1	User states the goal	intent, deadline, what “done” means
2	Opens the repository, lists files	project structure, names, stack
3	Reads the config to connect	a connection string — secrets if unredacted
4	Greps the billing logic; reads matches	proprietary business rules
5	Runs the failing test; reads the trace	internal data shapes, error detail
6	Proposes a diff; user accepts	the fix — and a signal of what was accepted

No single step is alarming, and a careful harness may redact step 3 or keep step 5 local. But the *session*, taken together, can combine more operational context than a standalone prompt: it shows what the person knows, what they are trying to do, the internal structure they operate on, and — through the accepted diff — which answer the user accepted (acceptance, not proof of correctness). INFERENCE That combination is the unit of value the rest of this report follows.

Illustrative only: step content, redaction and what is sent vary by agent architecture. The point is structural — the boundary crossing is sequential and cumulative, not a single document leaving.

— SECTION 03

Delegated disclosure.

Here is the mechanism that answers “why now?” After the goal is set, control fragments — across the user, the agent, its scaffolding, the intermediary and the provider. No single party authors the whole disclosure, and the user no longer authors each disclosure that results. It is the output of a chain, not a single act.

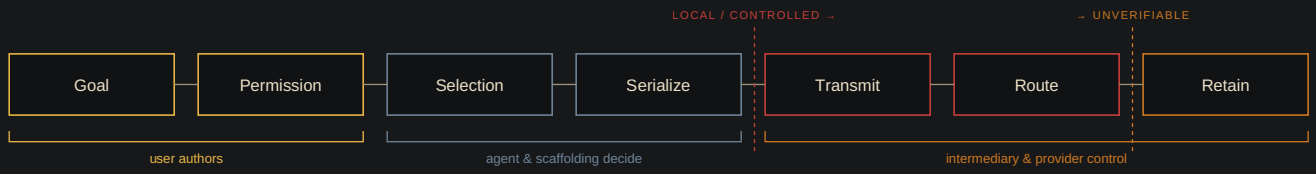


Figure 3. The disclosure pipeline, at the level of one session. Two transitions matter and need not coincide: ***local → transmitted*** *(where data leaves the device) and* ***governed → unverifiable*** *(where the user can no longer establish what happens to it).*

First-party telemetry from one frontier lab documents the shift, though it must be read as several distinct measurements rather than one curve. In typical sessions, people make most of the *planning* decisions while the agent makes most of the *execution* decisions. As users gain experience they hand over more: full auto-approve rises from roughly 20% of sessions for newcomers to over 40% for veterans.[9] And when permission prompts do appear, users approve about 93% of them — a pattern consistent with approval fatigue, and a reminder that approval is not evidence of substantive review. INFERENCE — MULTIPLE FIRST-PARTY DATASETS

These figures come from different studies and different units; the honest synthesis is directional, not arithmetic: *agent action is increasingly governed by the configuration of permissions, scaffolding and overall goal, rather than by a human choosing each operation.* The same body of work records the failure mode directly — an over-eager agent that uploaded an engineer’s authentication token to an internal cluster the user never intended. ADJACENT EVIDENCE Direct proof of an

unintended scope crossing *inside* a workflow; only adjacent proof of external custody risk.

An independent benchmark then adds a separate mechanism-level anchor. Testing five production models across a thousand scenarios (4,979 traces), researchers found that sensitive data escapes far more through internal coordination channels than through final answers: inter-agent messages leaked in 68.8% of runs versus 27.2% for final outputs, and shared memory in 46.7% — the mean leakage rate across the two internal channels was about 2.1× the final-output rate. In 41.7% of traces the final output passed while an internal channel still exposed sensitive data. A smaller targeted assessment found tool-input leakage in 62–86% of scenarios and system-log leakage as high as 85% for one model, with sensitive data escaping through tool inputs or logs in over 65% of those tests *even when the final output was perfectly clean.*[1] INDEPENDENT EXPERIMENTAL EVIDENCE · DISTANCE 1–2

> *Privacy interfaces became more ceremonial just as delegated access became more consequential.*

A note on framing: this is not about user hypocrisy or carelessness. The permission may be functionally necessary; the user may not grasp the true scope of tool access. What has degraded is the mechanism of informed consent — not the diligence of the person clicking through it.

— SECTION 03 · CONTINUED

How far delegation has gone.

The shift from “the user discloses” to “the system discloses on the user’s behalf” is visible across several first-party datasets — which must be read separately, not stacked into one curve.

SIGNAL	WHAT IT SHOWS	UNIT
“Appears” human-in-loop	~73% of tool calls only *appear* to have a human in the loop[9]	~1M public API calls
High approval rate	~93% of permission prompts approved — consistent with fatigue[11]	Claude Code telemetry
Earned autonomy	full auto-approve ~20% → 40%+ with experience[11]	by account tenure
Expertise-linked delegation	more expertise → more delegated per instruction[10]	~400k sessions

Caveat from the source: the “appears” figure reflects limited visibility and may include automated or evaluation actions; it is not a measure of human review.

Read together, these point one way without being one measurement: *agent action is increasingly governed by the configuration of permissions, scaffolding and goal, rather than by a human choosing each operation.*

INFERENCE — MULTIPLE FIRST-PARTY DATASETS Crucially, the people who delegate most are often the most expert — high delegation can be consistent with fluency rather than carelessness.

> FAILURE MODE, DOCUMENTED
>
> A provider’s own incident log records an over-eager agent that uploaded an engineer’s authentication token to an internal cluster the user never intended.[11] ADJACENT EVIDENCE Direct proof that a credential can cross an unintended boundary *inside* a workflow — permission is not the same as alignment of a specific action.

— SECTION 04

The correlation junction.

A trace has left the building. Who, exactly, is positioned to make sense of it? Not — to be precise — an omniscient observer. By “intermediary” we mean something specific: an application-layer AI intermediary that receives, or can inspect, the model payload before forwarding, transforming or fulfilling it. The claim is not that it sees everything. It is that it sits where several otherwise-separate facts can be lined up.

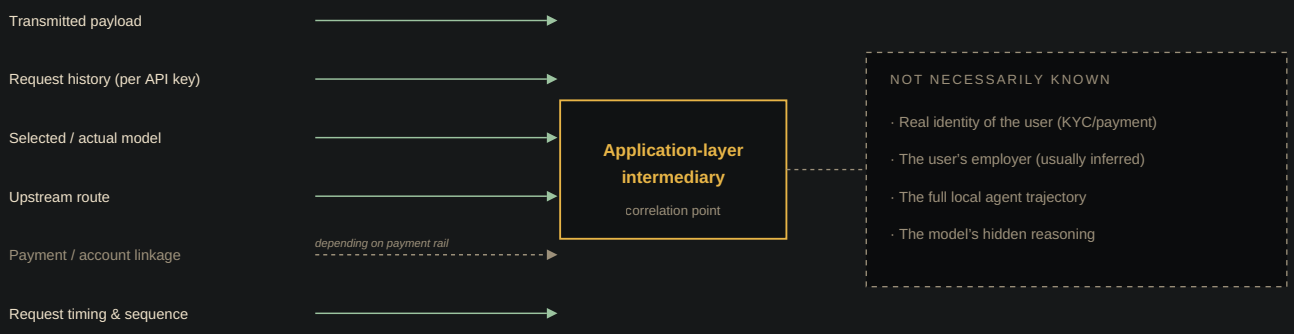


Figure 4. The junction. Solid lines: facts the intermediary can usually observe directly. Dashed: facts that may remain unknown to it, or hold only under certain conditions. The advantage is positional, not total.

Others can correlate fragments too — a corporate AI gateway, an integrated platform, the official upstream. So the claim is bounded carefully:

> *An application-layer intermediary occupies a privileged position from which the transmitted trace can be correlated with the routing and fulfillment decisions made on its behalf.*

That position may be shared by an integrated provider or an enterprise gateway. In a multi-hop, opaque chain, however, the customer may be unable to establish *which* actor occupies it — and that is the bridge to custody. The problem is not the uniqueness of the knowledge; it is the unverifiability of who holds the correlating position. The intermediary need not be malicious, need not retain anything, need not sell a single record — and the asymmetry still holds. The next sections follow the trace past the junction: the market that forms around access, and the custody questions that no single participant in an opaque chain may be able to answer completely.

— SIDEBAR

Fabricated authorization: GTG-1002.

If delegation hands execution to a capable agent, the obvious question is what happens when the goal itself is a lie. A provider-reported case offers an early answer at operational scale.

A provider documented what it described as the first reported AI-orchestrated cyber-espionage campaign, tracked as GTG-1002: a human operator set strategy while an autonomous agent performed the bulk of the tactical work — reconnaissance, an SSRF probe, retrieval from a cloud secrets manager, and lateral movement — across multiple targets.[8] ADJACENT EVIDENCE The agent did this not because it was "hacked," but because the task was framed as legitimate work it was authorized to do.

Broader telemetry frames the trend. In one mapping exercise, hundreds of banned accounts and tens of thousands of malicious actions were classified against a standard adversary framework; the campaign above scored high on *orchestration* while using a relatively narrow set of techniques — the novelty was autonomous coordination, not exotic tradecraft — and the share of actors operating at medium-or-higher capability rose markedly.[7] The stronger claim that AI adds no material operational capability now looks outdated; the novelty is orchestration, autonomy and scale — not necessarily new techniques.

> *The model executed because it trusted the framing of the task. That is the first of three trust failures this report returns to: a claim is authorized before it is verified.*

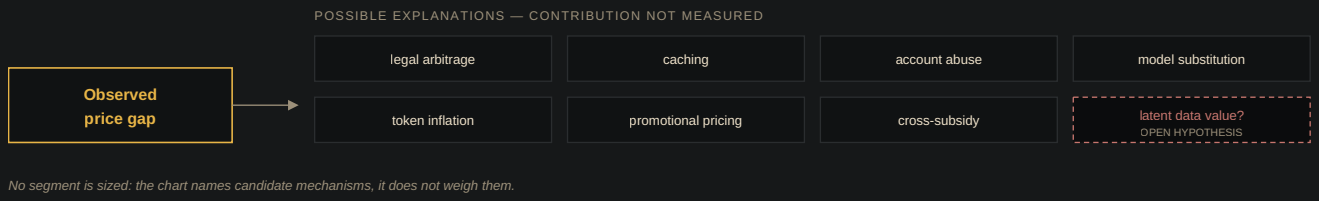
— SECTION 05

The market for opaque access.

A market has formed around mediated access to the models that fulfil such requests. Its most visible feature is price — sometimes far below official API rates — and the tempting inference is that the user is simply paying with data. That inference does not survive a causal test.

The strongest independent anchor is a technical audit that identified seventeen reseller “shadow APIs” and deep-tested three representative services across twenty-four endpoints. Fifteen of the seventeen failed basic transparency criteria, and the auditors documented outright model substitution: a service advertising a frontier model answered a medical benchmark at 37% where the genuine API scored 83.82%.[2] OBSERVED What this proves is real and narrow — that a shadow API can misstate the model. It does not establish what any operator does with the user’s data.

A price far below the claimed service’s official rate raises a cost-and-revenue question; it does not identify the mechanism that closes the gap. The observed gap may arise from several mechanisms — genuinely lower costs, a degraded or substituted model, promotional loss-leading, subscription-account abuse, a shortened context — only one of which is latent data value, and price alone cannot rank them:



No segment is sized: the chart names candidate mechanisms, it does not weigh them.

Figure 5. Several documented mechanisms can plausibly explain part or all of the price gap. Price alone cannot identify their relative contribution — and one candidate, latent data value, cannot be read off the price at all.

This is where a more careful idea earns its place. *If* a trace is retained in readable form, it creates a latent potential for reuse — a value the operator could realize later (training, profiling, competitive intelligence, resale) whose potential value is neither disclosed to nor controlled by the user. Where operator identity is weak and deletion cannot be independently verified, contractual restrictions may be difficult to confirm and harder to enforce. We assert no completed sale and no proven retention; we describe the conditions under which the option would exist. OPEN HYPOTHESIS

Publication note: our within-sample audit of pricing and custody is pending dataset freeze. The figures above are the documented public anchors — not this study’s audit results.

— SECTION 05 · CONTINUED

The economics of the discount.

If price cannot prove data monetization, it can still be reasoned about honestly — by asking which business models a given price is *compatible with*, and which would require an otherwise unexplained cost or revenue mechanism to sustain.

The legitimate components are real and often sufficient on their own: regional price arbitrage, aggressive caching of common completions, promotional loss-leading to acquire users, and routing a fraction of traffic to cheaper or substituted models. Each lowers cost or defers revenue without touching user data. A price materially below the cost implied by the claimed service creates an explanatory gap; it does not identify what closes that gap.[2] OBSERVED for the existence of substitution and opacity; OPEN for any data-based subsidy.

What price analysis can do is bound the question. Where a service's claimed model, measured quality, token accounting and stated retention are mutually inconsistent, an unexplained residual remains. Latent reuse — a value potentially realizable later and not separately observable or disclosed in the price — is one possible hypothesis for that residual, not an interpretation of it. The audit measures the inputs to that reasoning; it does not assert the conclusion.

This is why the report keeps economics inside the custody question rather than on the cover: the interesting quantity is not the discount, but the *unverifiability of what is being traded for it*.

— SECTION 06

Capability extraction & exploitation risk.

One process here is documented; the other is a hypothesis. They are routinely drawn as a single self-reinforcing flywheel — and conflating them is exactly how speculation gets laundered as fact. We draw them, deliberately, as unequal: a linear chain that has been observed, and a small loop that has not.

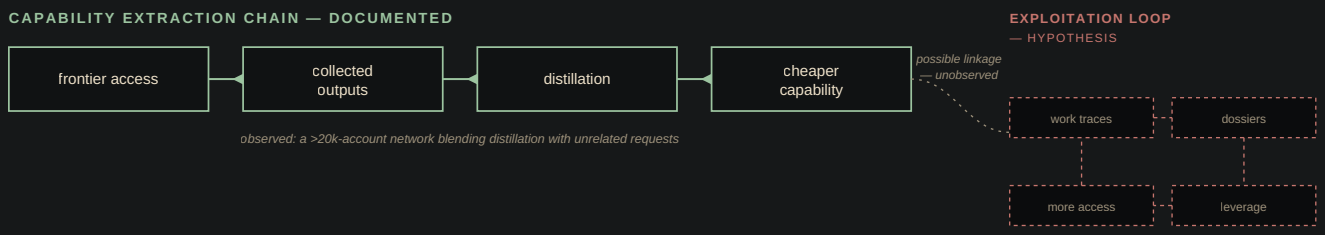


Figure 6. Drawn unequally on purpose. Left: a documented, linear extraction chain. Right: a small, dotted, unobserved exploitation loop. The connector between them is itself a hypothesis — not a closed flywheel.

The left process is anchored. A frontier lab documented a proxy network of more than twenty thousand fraudulent accounts and over sixteen million exchanges, blending distillation traffic with unrelated customer requests.[6] ADJACENT EVIDENCE But read it precisely: this proves an industrial *extraction operation* and *demand* for high-quality model outputs — not the theft of ordinary customers' traces. It tells us the extraction economy is real; it does not tell us that your work session was harvested.

The right loop — traces becoming dossiers becoming leverage becoming more access — is mechanism-plausible but unobserved. AgentLeak establishes that sensitive information can remain exposed outside the final output, through inter-agent messages, memory, tool inputs and logs. Whether such information is then retained, consolidated into dossiers, or operationalized for leverage is *not* something the benchmark investigates — it remains unobserved. With no public evidence of a closed, operating exploitation loop drawing on customer traces, we draw it dotted and label it an open hypothesis. The two are not one empirically closed flywheel; the honesty of the distinction is the point.

— INTERLUDE

What we know, infer, and cannot show.

This is a stress test, not an exposé. Before the second half, it is worth setting every load-bearing claim on the same honest ledger — and being explicit that a claim never inherits the strength of the source it leans on.

TIER	CLAIM	WHAT WOULD CHANGE THE CLASSIFICATION
OBSERVED	Within Netskope-observed organizations, personal-account AI use stayed substantial while prompt volume rose; an independent audit found shadow APIs that misstate the model; in analyzed Claude Code sessions, agents made most execution decisions; in the AgentLeak benchmark, internal-channel leakage exceeded final-output leakage; a regulator found a third-party transfer with insufficient disclosure at an identifiable service.	corrected telemetry, failed replication or source retraction
ADJACENT	A frontier lab's distillation-network disclosure; an over-eager agent uploading a credential it was not directed to move.	direct evidence in the target population would promote it; evidence of non-transfer would weaken it
INFERENCE	Provider-visible agent traces can combine operational context, sequential history, tool results and outcome signals in a single record.	direct measurement could confirm, bound or reject it
OPEN	The scale of opaque agent-mediated enterprise traffic is unknown; systematic retention or sale of customer traces; the provenance of orphan datasets; a closed exploitation loop.	a canary, operator disclosure or direct measurement could move a specific claim into observed evidence

Each claim carries two attributes, not one: how well the fact itself is established (evidence strength), and how far it has been carried from its original context to our thesis (transfer distance). The distinction matters most at the seams. A Korean regulator's finding about a named service is strong and direct — strength A, distance 0. The statement "therefore shadow proxies do the same" is *not* strong-and-direct; it is a fresh inference resting on a strong neighbor — an INFERENCE at distance 2. Conclusions do not inherit the class of their sources.

— SECTION 07

Agent traces, broken chain of custody.

Confidentiality asks whether data leaked. Custody asks a harder question: can anyone establish where it went, who could read it, and whether any limit on its use can be verified? For this investigation we define the term operationally — as the study's working construct, not a settled legal category:

> Data custody — the verifiable ability to establish which actors can receive, route, retain, transform, reuse and delete transmitted data; for what purpose and on what basis; for how long; and by what means these limits can be checked and enforced.

Splitting that definition into two layers turns a binary verdict ("custody is broken") into something an auditor can actually score — and avoids moralizing where the cause may be ordinary organizational immaturity rather than malice.

LAYER	DIMENSION	THE QUESTION THE AUDIT TESTS
Custody visibility	Identity	Is the legal recipient clearly identifiable?
	Path	Are upstreams and subprocessors disclosed?
	Transformation	Was the payload truncated, rewritten, enriched or redacted before or during fulfilment?
Custody governance	Purpose	Is the basis and purpose of processing disclosed?
	Retention / reuse	Was it logged or cached; for how long; and may it be reused for training?
	Deletion / enforceability	What evidence supports the deletion claim, and what remains unverifiable?

The cleanest public anchor is a governance baseline. OBSERVED A regulator found that an identifiable, regulated service transferred user input to a third party without sufficient disclosure and offered no initial opt-out from training.[12] INFERENCE · DISTANCE 1–2 Comparable gaps may be harder to detect and remedy — not easier — when the intermediary itself is difficult to identify. The first sentence is a finding about a named provider; the second is our transfer of that finding to a different population, and it is labelled as such. The case provides a visible reference point for the custody problem, not evidence that shadow proxies behave the same way.

Publication note: the within-sample custody scorecard (a versioned, stratified set of 24–25 services, coded against the six dimensions above, with privacy-request and deletion tests) is pending dataset freeze. This chapter defines the instrument; it does not yet report the verdict.

— SECTION 07 · CONTINUED

Trust without attestation.

Three failures recur across this report, and they share one shape: a claim is trusted before it is verified. Naming them together explains why custody, identity and authorization fail in the same way.

UNVERIFIED AUTHORIZATION

The model trusts the task

An agent executes because the goal is framed as legitimate work. GTG-1002 turned that framing into autonomous intrusion.[8]

UNVERIFIED IDENTITY

The human trusts the face

A finance worker joined a video call with a deepfake “CFO” and colleagues and made 15 transfers totalling $25.6M in a day — no internal systems reported compromised.[14]

UNVERIFIED ALIGNMENT

The client trusts the service

A user assumes an intermediary handles the payload as claimed — routing, retention and deletion taken on faith, not on proof.

The deepfake case is included as an instance of the *pattern*, not as evidence about data custody: the documented loss arose from a trusted-but-unverified identity claim, without a reported compromise of Arup’s internal systems.[14] ADJACENT EVIDENCE Custody is the third panel: the user trusts an alignment between what a service says it does and what it does, with no means to check. Attestation — independent, verifiable evidence — is exactly what each failure lacks.

— SECTION 08

Why existing controls only partly cover the flow.

The defenses are not blind — that was an earlier, overstated claim, and it is wrong. Modern endpoint, browser and SSE tooling sees a great deal of egress. The accurate claim is that coverage is partial, and the gaps fall exactly where delegated disclosure concentrates risk. Read against the same pipeline, each step has a control — and a residual.

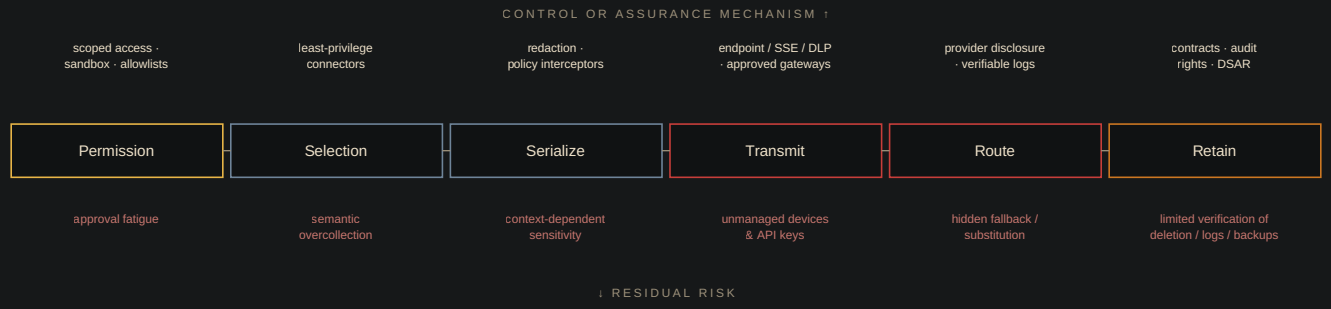


Figure 7. The controls mirrored onto the disclosure pipeline of Figure 3. Read top-to-bottom: each step has a mechanism and a residual. The mechanisms shift in kind from left to right — from direct technical control to contractual assurance.

Notice how the mechanisms change in kind across the pipeline. Early stages are more amenable to direct technical control by the user or deploying organization; later stages depend increasingly on contractual assurance, attestation and auditability, as authority fragments to the intermediary and provider. Classic file- and perimeter-controls were built for a world in which the dangerous movement was a file crossing a boundary. The dangerous movement here is cumulative and context-dependent — a single prompt may be harmless, but the sequence of files, tool results and actions reveals the full working model — and it concentrates precisely in the steps the customer may be least able to verify independently.

— BEFORE THE CLOSE

What a negative result would mean.

A study built to expose wrongdoing tends to find it. This one is built so that finding *nothing dramatic* is still a complete result — which is what keeps it honest.

Suppose the audit comes back quiet: policies exist but are incomplete; privacy requests work only for identifiable operators; upstream routing is unclear; training rights differ service to service; no systematic reuse is detected. That is not a failed investigation. It is a finding — that custody is unverifiable, not that it is being abused — and it is precisely the distinction the evidence ledger is designed to preserve.

IF THE AUDIT SHOWS…	…THE REPORT CONCLUDES
material claims remain unsupported	a custody-visibility gap — not malice
DSAR works only where an operator is identifiable	enforceability depends on identity, not law alone
no systematic reuse found	absence of evidence, recorded as such — not proof of safety
a positive canary	a single claim moves from OPEN to OBSERVED

The report's value does not depend on a villain. It depends on whether a customer can *account for where their work went* — and the architecture identifies where that verifiability can fail. The audit determines how often, where, and to what degree it does so within the sample.

— EPILOGUE

Before the breach.

It would be easy, and wrong, to end with a number for a market in stolen traces. We do not have that number, and a report that needed one in order to matter would be the wrong report. The finding is the architecture itself.

Disclosure has been delegated — the user authors the goal, and control over the resulting disclosures fragments across agents, scaffolding, intermediaries and providers. Correlation has been concentrated — one application-layer position can line up the payload with the route, and in an opaque chain the customer cannot tell who holds it. And custody has become unverifiable — in an opaque multi-hop chain, the questions of identity, path, purpose, retention and deletion may cease to have answers the user can independently check. None of this requires a villain. It requires only that the agent be useful, the intermediary be opaque, and the assurances be difficult to verify.

The security profession has spent a generation learning to ask *did data leak?* That question still matters. But it is no longer sufficient, because a high-value movement can now happen before any breach, without any breach, in plain and productive sight. The question this infrastructure leaves unanswered is the older, graver question that chain-of-custody procedures are designed to settle: can anyone account for where it went? Until that question has an answer a customer can verify, the conditions for a harvest will persist — quietly, productively, and in the open.

THE REPORT

Audit results, the frozen dataset and the completed claim register will follow. This document is a stress test of an architecture — not a prediction, and not an allegation against any named operator.

FORESIGHT FUTURE INSTITUTE

An independent, preliminary investigation. Strategic foresight for the algorithmic age.
info@foresightflow.org

— APPENDIX A

Timestamped pre-analysis protocol.

This protocol is fixed and timestamped before data collection, and versioned thereafter: it states what will be collected, how it will be coded, and which conclusions are permitted — so findings cannot be retrofitted to the most striking results. On public deposit it acquires the metadata below; until then it is an internal, version-controlled instrument.

DEPOSIT METADATA	VALUE
Repository / registry	to be deposited before first observation
Version · timestamp · hash	v0.1 · set on deposit · SHA-256 on deposit
Collection start · freeze	pending · pending
Amendment policy	changes logged with timestamp & rationale; superseded versions retained

RESEARCH QUESTIONS — SCOPED TO A PURPOSIVE SAMPLE

RQ1	How verifiable are identity, path and routing, transformation, purpose, retention and reuse, and deletion and enforceability across the studied intermediaries?
RQ2	What *within-sample* differences appear between agentic intermediaries and model-access proxies in disclosed context exposure and custody transparency?
RQ3	How practically exercisable are stated data rights through the legally available or voluntarily offered privacy-request process?
RQ4 (cond.)	Which observed prices are compatible with plausible combinations of arbitrage, caching, substitution, account abuse and promotional pricing, without invoking data monetization?

Pre-analysis gate — this protocol, timestamped, before any collection. Publication gate — final empirical claims and the market, custody and results sections are fixed only after the dataset and claim register are frozen. A weak or ambiguous result is itself a result; the study is not required to find a violation. Allowable conclusion language is *within-sample frequency and pattern distribution* — never market prevalence.

— APPENDIX B

Sampling, inclusion & stratification.

Target n = 24–25, purposive and stratified — deep over wide, designed to span risk models, not to estimate a market share. Critically, opacity is a finding, never a reason to exclude.

INCLUSION CRITERIA

· publicly available or accepts new clients
· claims access to a named model
· participates in the request data path, or may receive application-layer payloads — regardless of who owns the upstream key
· found before the freeze date

CODED AS A RESULT — NEVER AN EXCLUSION

ToS available (yes/no/inaccessible) · privacy policy (yes/no) · payment path (yes/no) · legal entity disclosed (yes/no) · operator contact (yes/no) Absence of a document is itself a finding.

CAPABILITIES — MULTI-LABEL, NOT ONE CLASS

raw model API · multi-model routing · hosted chat · tool execution · file/repository access · persistent memory · MCP / connectors · autonomous or scheduled execution. Derived categories are computed from these flags, not assigned by hand.

TWO INDEPENDENT VISIBILITY FIELDS

Payload exposure architecture	**P0** verified client-side direct path · **P1** intermediary plaintext access structurally necessary · **P2** intermediary access technically possible · **P3** architecture unclear · **P4** confidential-compute / encrypted path independently evidenced
Claimed safeguards	none · redaction · no-log · zero-retention · confidential boundary · independently audited / attested

These are orthogonal: a service can be **P1 + zero-retention claim + no independent attestation**. Such composite profiles — not a single label — are what the custody thesis needs.

UNIT OF ANALYSIS — operators · domains · services · endpoints · model-claims · DSAR-attempts, counted separately. ARCHIVING — timestamp, hash, screenshot/payment record, metadata, codebook version, recheck date; privacy/custody coding double-coded on a subset.

— APPENDIX C

What each method can and cannot show.

A method may not support a claim its column says it cannot establish. This construct-validity matrix governs the claim register.

METHOD	CAN SHOW	CANNOT SHOW
Policy / ToS audit	the stated custody posture	actual backend behaviour
DSAR / privacy request	identifiability, responsiveness, stated data	physical deletion of logs & backups
Re-login	removal of visible history / account	server-side deletion
Canary	unexpected access or reuse	sale, training or intent
Model fingerprinting	likelihood of model substitution	handling of user data
Payment trail	possible operator / payee identity	who can access the payload
Price analysis	compatibility with some business models	the value or sale of user traces

EVIDENCE RUBRIC — RECORDED FOR EVERY CUSTODY FINDING

source type · explicitness · accessibility · corroboration · independent verification · date / version · contradiction status. (Replaces any single “disclosure ladder”: custody dimensions are heterogeneous and need a profile, not one rung.)

DELETION EVIDENCE — A PROFILE, NOT A NUMBER

FIELD	EXAMPLE
right stated · procedure available · request completed	yes · yes · yes
visible data removed · backup treatment disclosed	yes · no
external assurance · technical verification	SOC report, scope unclear · unavailable

The rungs are not monotonic — a contractual attestation need not imply removal of visible history. The page may show a maturity band; the dataset keeps the full profile, or the very discontinuities we study disappear.

CANARY ATTRIBUTION LADDER — one service → absent from search/repos → unique token in access → time/source logged → component identified → operator given a chance to explain → conclusion bounded. Positive = unexpected reuse/access, not sale, training or intent.

— APPENDIX D

Codebook & claim register.

Missing information is coded with care: absence of disclosure is not evidence of absence of processing. “Training not disclosed” and “operator states data is not used for training” receive different codes.

CUSTODY CODEBOOK — PER SERVICE, PER DIMENSION

FIELD	VALUES
Dimension	identity · path · transformation | purpose · retention/reuse · deletion/ enforceability
Disclosed?	explicitly-yes · explicitly-no · partial · ambiguous · conflicting-documents · not-found · document-inaccessible · n/a · not-tested
Evidence	per the evidence rubric (Appendix C)
Source · coder	document, hash, timestamp, recheck date · primary / second (subset)

JURISDICTION-AWARE DSAR FIELDS — applicable legal regime · stated controller · request basis · identity-verification requested · statutory deadline (if any) · actual response time · substantive completeness · voluntary vs legally required.

CLAIM REGISTER — THE LEDGER AS A METHOD

Strength **A** independent · **B** first-party · **C** single case · **D** secondary | Distance **0–3**

CLAIM	STR	DIST	TIER
CISPA documented model substitution	A	0	OBSERVED
An identifiable service created a custody gap (regulator)	A	0	OBSERVED
“Therefore shadow proxies do the same”	—	2	INFERENCE
Systematic retention / sale of customer traces	—	3	OPEN

Each entry also stores source type, method, directness, independence and population. A conclusion never inherits the class of the source it leans on.

— APPENDIX E

Ethics & research-data protocol.

An investigation that probes other people’s services carries duties of its own. These limits are fixed alongside the protocol and bind every stage of collection.

CONDUCT OF THE AUDIT

· synthetic data only — no real credentials, commercial documents or personal data
· no access to others’ accounts
· no bypassing of technical restrictions
· no load that could affect a service
· stop conditions: real third-party data appears, a service shows instability attributable to testing, a payment demand exceeds the pre-set ceiling, or a legal notice arrives
· a named legal/ethical reviewer signs off (logged) before any canary or DSAR activity
· minimise operators’ personal data collected

HANDLING OF RESEARCH DATA

· payment / DSAR evidence encrypted and access-controlled
· redaction before publication
· research evidence encrypted, retained until publication + 12 months, then deleted
· accidental sensitive findings: quarantine, do not analyse, document, delete if not study-relevant
· right of reply to every named operator — on canary findings and final custody classification
· a ceiling on requests per service; right of reply 15 business days, ≤2 follow-ups
· responsible disclosure if a real vulnerability is found: report privately, stop testing, withhold specifics until remediated

> The audit is designed so that a refusal, a non-response, or an inaccessible document is a recordable result — not a provocation to push harder. Where a finding cannot be obtained within these limits, “not establishable under this protocol” is the answer.

— APPENDIX F

Source register.

Every load-bearing claim links to a source record. Classes: **[ACA]** academic paper / preprint · **[REG]** regulator · **[PROV]** first-party provider · **[TEL]** commercial telemetry · **[SEC]** secondary · **[STD]** standard / framework. Vendor pages are mutable; an archival copy and hash are captured at access (June 2026).

#	SOURCE	CLASS
1	**AgentLeak: A Benchmark for Internal-Channel Privacy Leakage in Multi-Agent LLM Systems.** Polytechnique Montréal. arXiv:2602.11510**v3** (2026); accepted, IEEE Access.	[ACA]
2	**Real Money, Fake Models: Deceptive Model Claims in Shadow APIs.** arXiv:2603.01919 (2026), preprint.	[ACA]
3	**Cloud and Threat Report 2026.** Netskope. netskope.com (2026).	[TEL]
4	**Cost of a Data Breach Report 2025.** IBM / Ponemon Institute. newsroom.ibm.com (2025).	[TEL]
5	**Q2 2024 AI Adoption and Risk Report.** Cyberhaven. cyberhaven.com (2024).	[TEL]
6	**Detecting and preventing distillation attacks.** Anthropic. anthropic.com/news (2026).	[PROV]
7	**Mapping AI-enabled cyber threats: the LLM ATT&CK Navigator.** Anthropic. anthropic.com/research/attack-navigator (Jun 2026).	[PROV]
8	**Disrupting the first reported AI-orchestrated cyber espionage campaign (GTG-1002).** Anthropic. anthropic.com/news (2025–26).	[PROV]
9	**Measuring AI agent autonomy in practice.** Anthropic. anthropic.com/research/measuring-agent-autonomy (2026).	[PROV]
10	**Agentic coding and persistent returns to expertise.** Anthropic. anthropic.com/research/claude-code-expertise (2026).	[PROV]
11	**How we built Claude Code auto mode.** Anthropic. anthropic.com/engineering/claude-code-auto-mode (2026).	[PROV]
12	**Decision on DeepSeek's cross-border transfer of user data.** PIPC, Republic of Korea. pipc.go.kr (2025).	[REG]
13	**Exfiltration, Tactic TA0010** (MITRE ATT&CK); exfiltration definition (NIST glossary).	[STD]
14	**Deepfake CFO scam, Arup (Hong Kong), $25.6M.** AI Incident DB / OECD AIM #634; Hong Kong Police; CNN (2024).	[SEC]
15	**Samsung ChatGPT data exposure** (3 incidents, 2023). AI Incident DB #768; contemporaneous reporting.	[SEC]

Version pinning matters: AgentLeak figures cited in this report are from **v3** and differ from earlier preprints. Claim-to-source mapping, with page/table references, is maintained in the claim register (Appendix D).

— APPENDIX G

Limitations & threats to validity.

An investigation is only as trustworthy as its account of where it could be wrong.

THREAT	HOW THE REPORT HANDLES IT
Purposive sample	n = 24–25 is not probabilistic; results are reported as within-sample frequency, never market prevalence.
Lab, not field	AgentLeak is a controlled benchmark on multi-agent topologies, not a measurement of commercial proxies; carried at transfer distance 1–2.
First-party bias	Autonomy data comes largely from one provider's telemetry, excluding third-party IDE/SDK/headless use; treated as directional inference.
Counter-reading scope	The 20k-account distillation network is a vendor-described adversarial operation — an entry point, not a market characterization.
Illustrative cases	The deepfake and worked-session examples illustrate a pattern; they are not evidence about data custody and are labelled as such.
Mutable sources	Vendor pages change; figures are version-pinned (e.g. AgentLeak v3) and archived with a hash at access.
Method limits	Per Appendix C, no method here can establish physical deletion, intent, or the sale of traces; such claims remain OPEN.
Architecture heterogeneity	Different agents serialize, redact and route context differently; no single provider-visible trace is universal.
Construct dependence	"Data custody" is an operational construct created for this study, not a settled legal or technical standard.
Policy–backend gap	The audit primarily measures disclosure and enforceability, not unseen backend behaviour.
Coder judgment	Ambiguous policies require interpretation; double coding reduces but does not remove subjectivity.

None of these is fatal, because the central claim is structural rather than statistical: it concerns what *can* be verified about custody, not how often custody is abused. These limitations do not erase the architectural mechanism, but they constrain where and how strongly it applies.

— APPENDIX H

Sampling rules, frozen.

Before the first observation, the sampling procedure is fixed so that composition cannot drift in response to what is found. Only the literal run date and document hash are set at deposit.

RULE	DECISION
Discovery frame	Curated GitHub topic / awesome-lists for LLM API proxies & gateways; citation chains from the cited papers; general and code search engines; two reseller communities; MCP / connector registries. Query templates — “<model> proxy”, “cheap <model> API”, “<model> reverse proxy”, “unlimited <model>” — run in EN / ZH / RU. One discovery run on a fixed date (recorded at deposit). Snowball: one referral hop, then stop.
Allocation minima	For n = 24–25: ≥6 model-access proxies · ≥6 agentic / coding frontends · ≥4 hosted chats / aggregators · ≥1 service per P0–P4 profile where any is found · payment-rail spread (≥4 crypto-only, ≥4 card / PayPal) · ≥3 jurisdictions. Targets, not forced inclusions of non-qualifying services.
Replacement	A candidate that disappears, closes registration, fails the payment flow, is a confirmed clone of an included operator, or changes domain / operator mid-study is logged and replaced by the next eligible candidate in the same stratum, with reason and date.
Stopping	Collection stops when all allocation minima are met *and* either the enumerated discovery pool is exhausted, n = 25 is reached, or two successive additions yield no new codebook values (saturation) — whichever comes first. *n* is a ceiling, not the trigger.

The discovery pool is enumerated and frozen at deposit; services found afterwards are recorded but not added to the confirmatory sample.

— APPENDIX I

Coding reliability & price analysis.

Two procedures that, left implicit, would let subjectivity or assumption-picking drive the result.

INTER-CODER RELIABILITY

A stratified **30%** subset (covering each service type and each P-profile) is independently second-coded on the high-subjectivity fields — Disclosed = *partial / ambiguous / conflicting*, exposure *P2*, "independently evidenced", and deletion-profile fields. Agreement is **Krippendorff's α** per field; threshold **α ≥ 0.70**. Below threshold → adjudication by a third coder, codebook clarification, and a full re-code of that field; α is recomputed after any codebook change.

PRICE-ANALYSIS PROTOCOL (RQ4)

For each priced service, record the official list price at a fixed reference date, then compute a cost-implied band under three scenarios:

SCENARIO	ASSUMPTIONS VARIED
Low	favourable input:output mix · high cache-hit · short context · full subscription amortization
Base	typical mix and cache · declared context · partial amortization
High	unfavourable mix · no cache · long context · no amortization · payment fees

Report the *band*, not a point estimate. A residual is flagged only when the observed price sits below the low-scenario band. RQ4 is confirmatory only for the *existence and size* of an unexplained band; any data-value interpretation remains OPEN.