ClearThought.ai

INSIGHTS · WHITEPAPER · OCTOBER 2026

Cost recovery stalls at the *access layer.*

What public-sector assessments teach about timelines, price truth, and where AI belongs in cost recovery.

10–12 wk

To a priced baseline when access is approved up front

3 prices

List, contracted and billed, reconciled on every finding

1M+

Contract pages our principals processed under governed, logged inference

READ ONLINE AT CLEARTHOUGHT.AI/INSIGHTS/ACCESS-LAYER/

In short. Public estates are rarely hard to analyze; they are hard to *see.* Most of an assessment's calendar goes to waiting for read access, working out which price the organization actually pays, and telling refused reads apart from clean results. Fix the access path and the price basis first, run deterministic and AI-assisted evaluation side by side, and govern where inference runs. The same scope then lands in about half the time. Cost recovery here means finding and recovering overspend: seats no one uses, capacity paid for twice, and billing above contract. Recovering shared IT costs from federal awards draws on the same evidence and is covered on our Government page.

WHAT WE SEE

Estates are readable. *Access is the hard part.*

Across the public-sector estates we assess (cloud tenants, virtualization, licensing and contracts, often spread over several departments and regulated programs), the technology is rarely the obstacle. The same handful of patterns decides how fast an assessment moves and how much it finds.

WHAT WE SEE	WHAT IT COSTS	HOW WE FIX IT
Read-only access treated as a new request for every platform and reviewer.	Weeks of calendar time before any analysis starts.	One pre-reviewed access packet, approved before kickoff.
Refused or narrowed reads that look like clean results, and permissions that drift between scans.	Savings undercounted; risks reported as absent.	Every refused read reported as a named gap, and scope checked on every rerun.
Savings priced at list instead of what the organization actually pays.	Figures on the wrong base, and credibility lost in review.	Contract, list and billed prices reconciled before anything is counted.
Licenses that no one reconciles against what is deployed or bought.	Paying for capacity twice, or for seats no one uses.	License counts traced from contract to tenant to user.
Tags that exist in name only, with placeholder values.	Shared cost cannot follow the programs it serves.	A short, enforced key standard and a measured baseline.

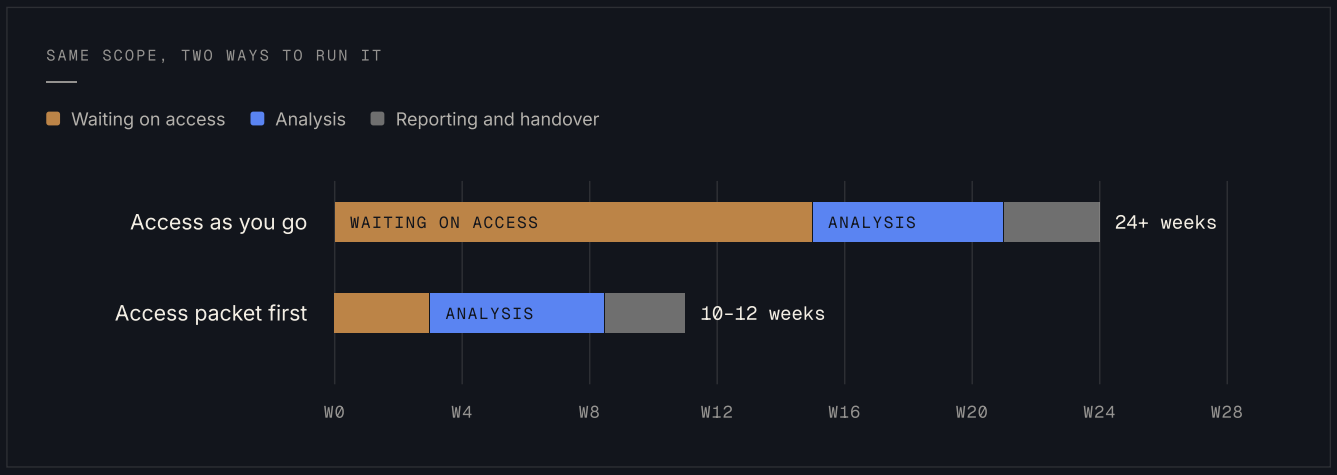
WHAT GOOD LOOKS LIKE

Assessments move at the speed of approvals. The fastest ones *name one owner for access* and put a clock on every request, so reviewers ask their questions in parallel rather than in sequence.

HOW WE RUN IT

Access first, *then everything moves.*

When access is requested platform by platform, most of an assessment's calendar is spent waiting. Once data arrives, scanning and repricing take hours, not weeks. Front-loading access changes the shape of the whole engagement.



Illustrative. The analysis itself barely changes; the waiting does.

What we put in place

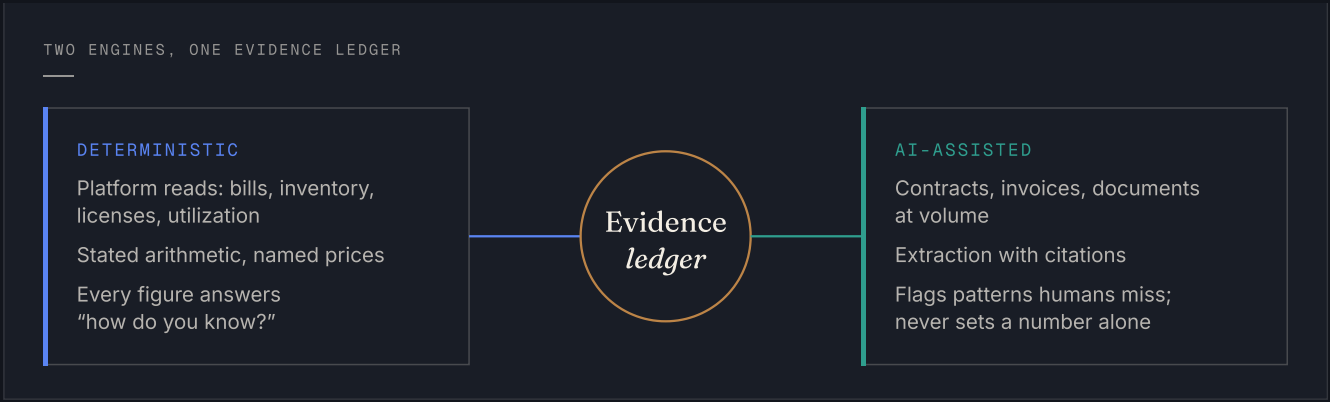
- **A pre-approved access packet before kickoff.** One reviewed bundle per platform: read-only roles as scripts, every permission listed with what it feeds, signed off by security before day one.
- **One accountable approver per department, with a 48-hour clock** and a named escalation path the sponsor enforces.
- **Read-only work exempted from change freezes in policy,** and freeze windows mapped at kickoff, not discovered mid-project.
- **Price truth in week one:** contract price sheets, renewal and reduction deadlines, and how each cloud estate is actually billed.
- **A weekly access burndown** in front of the sponsor, plus an automatic check that every scan sees the scope it saw last time.
- **Acceptance criteria fixed up front,** with data deliverables accepted separately from narrative ones.

COST RECOVERY

Deterministic first, *AI where it earns its place.*

Our internal proprietary tooling runs two engines against the same evidence ledger. The deterministic engine reads what platforms report (bills, inventory, licenses, utilization) through read-only collectors, and prices every finding from one named price list. Each figure has to answer “how do you know?” with platform data or stated arithmetic; anything inferred from a naming pattern stays out of the totals.

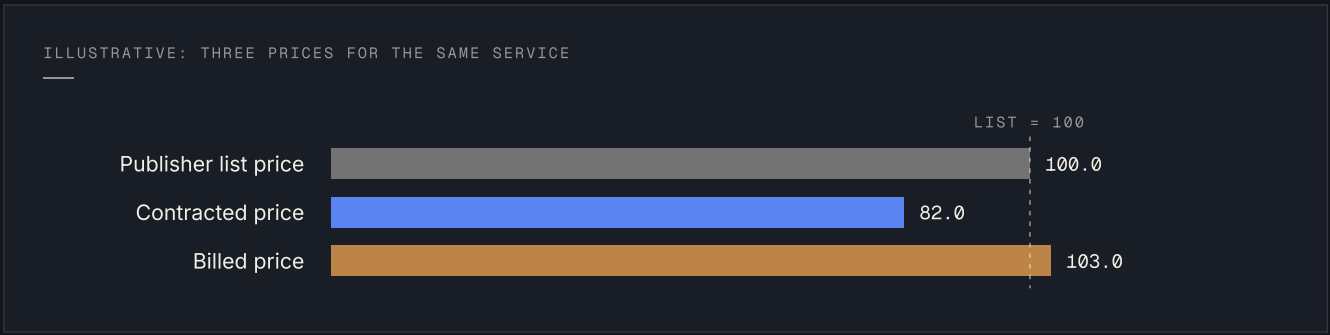
The AI engine does what people cannot do at volume: it reads contracts, orders and invoices, extracts prices, terms and renewal dates with citations to the page, and flags patterns for review. It never sets a savings number on its own. Its extractions feed the price list, and the deterministic engine checks them against what the platforms actually show.



The engines check each other: when a contract and a platform disagree, the result is a flagged question, not a saving.

Price truth changes the answer

List, contracted and billed prices are three different numbers, and savings move with whichever one you use. We regularly find the billed price differs from the contracted one, sometimes above list, when purchasing runs outside the agreement. Renewal dates matter as much as rates: license reductions usually have to be filed weeks or months before the renewal they apply to.



The FinOps Foundation's FOCUS specification draws similar lines with *ListCost*, *ContractedCost* and *BilledCost*, adds *EffectiveCost* to spread commitment purchases over the usage they cover, and uses separate invoice-issuer and service-provider fields that show when a reseller, not the provider, issues the invoice. We are aligning our internal model to FOCUS for cloud and SaaS billing, and will adopt it wherever it proves better in data and modeling. An internal data center can publish its charges in FOCUS form, but FOCUS does not say how to cost owned capacity on premises (hardware, depreciation, support, power and space), so that part of the hybrid model stays ours.

GOVERNANCE

Inference has to run *where you decided it would.*

AI-assisted cost recovery only works if leadership can say, for every document class, where the model ran and who could see the data. That is a control problem before it is a model problem.

- **Classify first.** Data that policy bars from AI processing never enters the pipeline; regulated classes are routed by rule, not by judgment.
- **Locally hosted clusters** for data that may not leave the building: models pinned to on-premises hardware, no outbound path, keys held by the agency.
- **In-cloud inference** in a dedicated account and region: models pinned by policy, no training on customer data, private endpoints, customer-managed keys, and deny rules that stop requests from routing elsewhere.
- **Every invocation logged and retained,** with throughput quotas planned in advance; at volume, throttling is a capacity decision, not an error.
- **People in the loop** for any extraction that changes a number.

DATA CLASS	WHERE INFERENCE RUNS	CONTROLS
Contracts, invoices, public records	In-cloud, dedicated account and region	Every call logged; models pinned by policy
Personal and staff data	In-cloud on private endpoints, or local	Customer-managed keys; access reviewed
Regulated (justice, tax, health, critical infrastructure)	Local cluster or government cloud, by rule	No outbound path; program owner approves
Barred by policy	Not processed	Excluded at intake, before any model sees it

1M+

CONTRACT PAGES

In prior work, our principals processed more than **one million contract pages** (about 55,000 documents) through a governed in-cloud pipeline in a dedicated account. Every model call and access was recorded: more than two million audit records over the life of the work.

HOW IT STARTS

Assess, architect, build, *hand over.*

PHASE	WHAT YOU GET
Assess	Access packet, read-only scans, a baseline priced on your own contract terms, and the recovery opportunity sized.
Architect	Target architecture, sequencing around renewals and freeze windows, and the funding streams worth pursuing.
Build	Delivered in phases, each with a rollback plan.
Hand over	Runbooks, standards and evidence your team owns and can rerun.

> Pricing follows the size and shape of the work. Depending on scope, we use fixed fees, time and materials, phased budgets, ongoing advisory arrangements, or a combination, matched to your procurement rules and funding source. We propose a structure once we have seen the system.

NEXT STEP

Bring us the system that *worries you.*

A 30-minute review with a principal, by video or at our offices in Houston, San Juan, or Christiansted. No sales engineers, no junior bench. We will tell you what read access the assessment needs, how long it should take, and where the recovery is likely to sit.

Book a review →

BRING TO THE REVIEW

- **Your estate list:** cloud tenants and subscriptions, virtualization platforms, and who administers each.
- **Your price truth:** contract price sheets, renewal and reduction deadlines, and how cloud spend is billed today.
- **Your approval path:** who signs off on read-only access, and any freeze windows ahead.

hello@clearthought.ai · clearthought.ai/contact

OFFICES

HOUSTON	1800 Saint James Place, Suite 218 Houston, TX 77056
SAN JUAN	619 Calle Ernesto Cerra San Juan, PR 00907
ST. CROIX	59 King's Wharf (King Christian Hotel Complex) Christiansted, VI 00820

Examples are illustrative and drawn from our public-sector work. Results depend on contract terms, renewal timing and access. Phases are sized to fit client requirements and realities.