THE TRUE COST OF BUILDING PRIVATE AI

What Regulated Businesses Need to Know

A decision framework for security leaders at regulated firms..

AT A GLANCE

The trap Shadow AI is accumulating exposure on vendor servers today. The build timeline is not a safe waiting period. It is the period of maximum risk.	**Build reality** 18 to 36 months. $1.5M to $2.5M before go-live. Hiring bottlenecks at every phase. And a $200,000/month opportunity cost meter running throughout.
The right question Is private AI infrastructure your core product? Or is the work your team does with AI your core product? If the latter, do not build it.	**Buy reality** Typical: 4 weeks to go-live. Evidence pack on day one. Training included. Your team working privately while the competition is still in budget approval. → totallyprivate.ai/roi

If your organization is considering building its own private AI platform, you have already identified something important: **your teams need AI, and public tools are not safe enough for your data.**

That insight is correct. The question this brief answers is whether building the infrastructure yourself is the best use of that capability, and what happens to your compliance posture in the meantime.

The firms that will lead AI adoption in regulated industries are not the ones that move fastest without governance. They are the ones that found a way to adopt AI safely without redirecting their core engineering capability to an infrastructure problem that has already been solved. The opportunity cost of delay runs at **$200,000 per month** for a 100-person team. That number is explained in full on page 4. It belongs in the first line of your build vs. buy analysis.

What the right answer looks like:

Private AI your teams actually adopt on week one, not quarter three.
Approval with evidence, not exceptions.
Audit artifacts on day one.
Training that turns access into hours saved.

> **Before you scope the project: is private AI infrastructure your core product, or is the work your team does with AI your core product? That question belongs at the beginning of the analysis, not the end.**

The Risk That Is Already Here

There is a version of the build conversation that assumes the status quo is neutral: your team is not yet using AI, you will build a safe platform, and then AI adoption begins. That assumption is almost certainly wrong.

In most regulated organizations, AI adoption has already started. It is happening informally, on personal accounts, on personal devices, outside any governance framework. This is Shadow AI, and in a regulated industry it is not a policy problem. It is an active liability.

We see the same pattern across regulated firms: AI use starts informally, then governance arrives after the first near-miss.

Delay does not stop AI adoption. **It just makes it undefended.**

If your industry is...	Shadow AI creates this specific exposure
Financial services	Client portfolio data, MNPI, deal flow, and trading strategies pasted into public AI tools. That data now lives on vendor servers subject to subpoena, regulatory examination, and breach. Regulators globally are actively scrutinizing AI vendor risk in financial institutions.
Healthcare	Patient names, diagnoses, and treatment details in AI prompts constitute unauthorized disclosure of protected health information. Depending on jurisdiction, a single incident may trigger mandatory breach notification. Patient health data is a specially protected category under most data protection frameworks globally.
Legal	Client matter details and legal strategy in a public AI tool may constitute a waiver of attorney-client privilege. Bar association confidentiality rules apply to every tool a lawyer touches, including AI.
All regulated industries	Some vendors may train on submitted data, and most create records on infrastructure you do not control, cannot independently verify or audit, and cannot confirm deletion from.

If you cannot answer **"Where does this data go?"** today, you already have an AI governance gap. This is the same question auditors and insurers are now asking.

> *Shadow AI does not pause while you build. Every month your team uses public AI tools is another month of exposure accumulating on vendor servers. The build timeline is not a safe waiting period. It is the period of maximum risk.*

A Security Property No Public AI Deployment Can Match

Totally Private AI (TPAI) includes a private API for agent integration, enabling automated AI workflows that process documents, summarize findings, draft communications, and interact with internal systems. This architecture creates a security property that on-premises and public AI deployments cannot replicate.

> ***Prompt injection exfiltration*** **is an attack where malicious instructions embedded in data processed by an AI agent attempt to redirect the agent to send sensitive information to an external endpoint. In an air-gapped environment with no public internet egress, there is no external endpoint. The attack vector is eliminated by architecture, not by detection, not by policy.**

This means your compliance team can approve AI agent workflows with a level of confidence that is architecturally grounded, not dependent on ongoing vigilance or vendor promises.

The Organizational Journey Before a Line of Code

A private AI infrastructure project at a regulated firm does not start with an engineering sprint. It starts with one person, the internal champion, building a business case and navigating an approval process that moves on annual budget cycles.

Phase	Realistic Timeline
Identify requirement, build initial business case	1 to 2 months
Build executive sponsorship and political alignment	2 to 4 months
Submit for annual budget approval	3 to 6 months. Miss the cycle: add 12 months.
Receive approved headcount and capital	Next budget cycle if deadline missed.
Begin hiring key technical roles	See next section.

If the initiative lands after the budget lock, the earliest a funded project begins is 14 to 18 months from the day the case was first made. During every one of those months, your teams are using public AI tools. The risk clock does not stop.

Risk accrues fastest before funding exists.

The Talent Problem in Regulated Industries

Once budget is approved, the project needs people. Specifically: a cloud architect with security experience, a security architect with AI and cloud familiarity, developers who understand compliance-driven design, QA with security testing background, and ongoing DevOps support. Finance, healthcare, and legal firms are not typically strong in these skills. That is not a criticism. It is a structural reality.

> *One search for a senior cloud architect at a large firm ran for 18 months. The candidate pool was consistently weak across repeated search cycles. The role was ultimately absorbed by existing staff whose primary responsibilities were already fully allocated. Security architect searches in the current market take longer.*

Role Required	Realistic Hiring Timeline
Cloud architect (security-focused)	6 to 18 months. Demand exceeds supply in regulated sectors.
Security architect (AI and cloud)	9 to 24 months. The hardest technical search in the current market.
Compliance specialist (redirect or hire)	2 to 4 months if redirecting existing staff from current duties.
Senior developers (2 to 3 FTE)	3 to 6 months each. Architecture must precede development.
QA with security testing background	2 to 4 months.
DevOps and operations (ongoing post-launch)	2 to 4 months. Required indefinitely after go-live.

Hiring runs in sequence, not in parallel. You cannot hire developers until the cloud architect has defined the architecture. You cannot finalize the architecture until the security architect has signed off on the design. This is not a failure of process. This is how regulated infrastructure gets built correctly.

"We Will Run It On Our Own Hardware"

Some organizations, typically targeting under 100 seats, choose to host AI infrastructure on-premises. For small deployments this can function, though gaps in availability, disaster recovery, auditing, and compliance evidence typically emerge during the first audit cycle. For enterprise deployments at scale in regulated industries, on-premises hosting introduces a category of problems that a server in a closet was never designed to solve.

Inference performance under concurrent load is one concern: without proper load balancing and queuing infrastructure, latency degrades significantly as simultaneous user sessions increase, and the capacity ceiling is fixed at hardware purchase with no ability to scale. But performance is the least of the problems.

Requirement	Server in a Closet	TPAI Cloud Deployment
Capacity and scale	Inference latency degrades under concurrent load. Capacity is fixed at hardware purchase. GPU servers cost $30K to $150K each.	Elastic compute via AWS Bedrock. Scales automatically with usage. No capital outlay.
High availability	Single point of failure. No SLA. Hardware failure takes the entire deployment offline with no timeline to restore.	Multi-AZ deployment. Automatic failover. 99.9%+ uptime.
Disaster recovery	Typically none. Data on failed hardware is a recovery project, not a procedure.	Automated backups. Point-in-time recovery. Cross-region option.

Requirement	Server in a Closet	TPAI Cloud Deployment
Authentication (SAML/SSO)	Custom implementation required. Enterprise identity integration is non-trivial to build and maintain correctly.	SAML 2.0 and OIDC included. Integrates with your existing identity provider.
Admin controls and user management	Custom build required: role-based access, seat provisioning, instant revocation, admin vs. user separation.	Centralized admin panel. Role-based access control. Instant user revocation.
LLM and software upgrades	Every model version and platform release requires manual testing, validation, and redeployment. Engineering time, every cycle, indefinitely.	Model updates automatic via Bedrock. Platform updates managed by Totally Private AI (TPAI).
Air-gapped security posture	Depends entirely on your physical and network controls. No independently verified posture. Air gap is a claim, not an architecture.	Logically air-gapped by architecture. No public internet egress. VPC endpoint isolation. SCPs prevent egress creation.
Defense in depth	Requires deliberate design: network segmentation, endpoint controls, application hardening, logging. Each layer must be built and maintained.	Multi-layer controls: network isolation, per-user encryption, operator access controls, immutable audit logs.
Legal holds and record retention	No built-in capability. Custom implementation required. Regulated industry standard: 7-year minimum retention.	Immutable audit logs. Configurable retention policies. Legal hold support. Evidence-ready exports.
Compliance audit trail	Custom logging. Tamper-evident audit trails are difficult to implement correctly and harder to maintain under audit scrutiny.	CloudTrail, VPC Flow Logs, application-level logs. Immutable. Evidence pack on request.
Physical security	Your server room. Your controls. Auditors will examine physical access logs, environmental controls, and change management.	AWS data center standards. SOC 2 and ISO 27001 certified facilities. Audit scope inherited.

> *A server in a closet is not a cost-saving option. It is a capital project with no SLA, no DR, no compliance audit trail, no legal hold capability, no enterprise authentication, and a single point of failure that takes your entire AI deployment offline when a hard drive fails at 2am.*

The Adoption Gap: Infrastructure Without Training

A private AI platform your team does not know how to use effectively is not a competitive advantage. It is an expensive tool sitting idle.

You can drop the most capable AI platform in the industry on someone's desk. Without training, they will use it the same way they use a search engine, and wonder why the results are mediocre. They will paste in a document and ask a vague question. They will get a generic answer. They will decide the tool is not as useful as they were promised, and it will stop being used within two weeks.

TPAI includes structured onboarding and workflow-specific training as part of deployment. Not a video library. Practical training built around the actual tasks your team will use AI to accelerate: documentation, research, client communications, compliance work. The difference between a user who prompts well and one who does not is measured in hours of recovered productivity per week.

> *Training is not an optional add-on. It is the difference between a platform that transforms how your team works and one that becomes a line item nobody defends at the next budget review.*

The True Cost Comparison

A realistic cost model for a 100-plus seat private AI deployment at a regulated firm includes labor, infrastructure, compliance, organizational runway, and the opportunity cost of the period between today and go-live. The figures below use conservative US market rates.

Note on AWS infrastructure costs: TPAI deploys into your AWS account. Your organization pays its AWS bill directly to Amazon, separately from TPAI's setup and maintenance fees. The figures in the Build column reflect what your team would spend building and managing equivalent infrastructure themselves.

Cost Item	Internal Build	TPAI Enterprise
Organizational runway (budget process)	$0 direct cost. 6 to 18 months of Shadow AI exposure accumulating daily.	Typical: 2 to 4 weeks to go-live for a 100 to 250-seat deployment. Shadow AI exposure ends on day one. You can scope the POC before you commit to Enterprise.
Cloud and security architect recruiting	$50K to $120K in search fees. 6 to 18 months elapsed.	Not required. Architecture delivered and maintained.
Development labor (12 to 18 months)	$600K to $1.2M. Architecture, development, security hardening, compliance mapping.	Not required.
Compliance review and framework mapping	$100K to $300K. External review plus internal effort per framework.	Evidence pack provided. Compliance mapping included.

Cost Item	Internal Build	TPAI Enterprise
AWS infrastructure (your AWS account)	$80K to $150K in Year 1 to build equivalent VPC, compute, logging, IAM, and SCP configuration.	Your AWS bill continues, but at operational cost, not build cost. TPAI setup fee covers architecture.
Ongoing maintenance (per year)	Fully-loaded cost of 0.5 to 1 FTE security and DevOps engineer: $150K to $300K. This is comparable to TPAI's annual maintenance fee for a similar seat count.	Annual maintenance fee. Architecture, patching, updates, incident response, and model governance included.
LLM upgrades and model governance	Engineering time every release cycle, indefinitely. Each new model version requires validation and redeployment.	Automatic via Bedrock. No engineering required.
Training and adoption program	Typically unbudgeted. Adoption failure is the most common enterprise software deployment outcome.	Workflow-specific onboarding included.
Total Year 1 build cost	$1.5M to $2.5M conservatively, before the system is live.	See the interactive ROI model: totallyprivate.ai/roi
Deployment timeline	18 to 36 months to production-ready in a regulated environment.	Weeks from contract to your team working privately.

The Opportunity Cost Calculation

The figures above cover direct costs. The larger number is the one that does not appear in most build analyses: the productivity your team does not recover while they wait for a safe AI platform.

For an interactive model you can adjust with your own headcount and rates: totallyprivate.ai/roi

Assumption	Value Used (Conservative)
Team members who would use AI	100 employees
Fully-loaded hourly cost per employee	$100 per hour (conservative for regulated industries)
Hours recovered per employee per week with AI	5 hours (documentation, research, compliance drafting, client communications)
Weekly productivity gain per employee	$100 x 5 hours = $500
Weekly gain across the full team	$500 x 100 employees = $50,000 per week
Monthly opportunity cost of delay	$50,000 x 4.3 weeks = $215,000 per month
Over an 18-month build cycle	$215,000 x 18 months = $3.87 million

> **$200,000 per month. Every month the project is in procurement, hiring, or development is another $200,000 in productivity your team did not recover. Over a realistic 18-month build cycle, the meter runs to nearly $4 million before a single user works privately.**

That number does not appear in most build vs. buy analyses. **Start with the meter.**

Run this model with your own headcount, hourly rate, and build timeline at totallyprivate.ai/roi. It includes a dedicated Build vs. Buy sheet.

What the Build Requires, Component by Component

Component	Internal Build	TPAI
Model gateway	Design, build, test, and maintain across every model version and update cycle.	Included. Multi-model via AWS Bedrock. Updates automatic.
Network isolation	SCPs, VPC egress rules, private endpoints, ongoing configuration audit.	Architecture-enforced. No public internet egress. Verified posture.
Per-user encryption	KMS integration, key management design, rotation policy, audit logging.	Included. Customer-managed keys.
Enterprise authentication	SAML 2.0 or OIDC integration with your identity provider. Custom build.	Included. SAML 2.0 and OIDC. Integrates with your existing IdP.
Audit logging and legal holds	Custom implementation. Tamper-evident trails are hard to build and harder to defend under audit.	Immutable logs. Legal hold support. 7-year retention configurable.
Compliance evidence generation	200 to 400 hours per framework. Frameworks vary by jurisdiction and sector; build teams typically address three to five independently.	Evidence pack included and maintained.
Agent and API integration	Custom build. Prompt injection attack surface must be addressed separately in design.	Private API included. Air-gapped architecture eliminates prompt injection exfiltration by design.
Training and onboarding	Typically unplanned. User adoption failure is the most common enterprise AI deployment outcome.	Workflow-specific training included.
Incident response	Internal. No dedicated SLA. Competes with every other engineering priority.	Founding team. One product. Direct access.

The Question Worth Asking First

> **Is private AI infrastructure your core product? Or is the work your team does with AI your core product?**

If you are a company whose competitive advantage is building and operating private AI infrastructure, build it. That is your business.

If you are a company whose competitive advantage is the legal work, the compliance decisions, the investment analysis, the patient care, the skilled professional judgment your team applies every day, and you need AI to do that work safely and compliantly, the build argument is an argument for creating a second infrastructure product inside your organization. One you will staff, maintain, defend to auditors, and evolve for years in a domain that is changing faster than almost any other.

TPAI deploys into your AWS account. You own the environment. You hold the keys. You control access. You set the retention policies. Your compliance team gets the evidence pack. Your users get trained. We run the infrastructure so your team does not have to.

That is not a compromise on control. **It is a decision about where your best people spend their time.**

See the architecture. Bring your security team.

30 minutes. We walk through exactly what is deployed, where it lives, what access controls exist, and how compliance evidence is generated. You leave with: data flow diagram, access model, and evidence list.

Book: calendly.com/rcastillo-totally-private

Email: rcastillo@totallyprivate.ai

TOTALLY PRIVATE AI

Your Data. Your AI. Totally Private.

totallyprivate.ai