



TOTALLY PRIVATE AI

---

# AWS Infrastructure Investment Guide

## Enterprise Deployment

*A transparent breakdown of your AWS infrastructure investment when deploying TPAI Enterprise. Every line item explained, every assumption stated, every optimisation opportunity identified.*

March 2026 | Version 2.0 | Confidential

# Contents

|                                                               |           |
|---------------------------------------------------------------|-----------|
| What Drives Cost Variance.....                                | 4         |
| What You Control.....                                         | 4         |
| <b>1. Overview.....</b>                                       | <b>5</b>  |
| Deployment Model.....                                         | 5         |
| Key Assumptions.....                                          | 5         |
| <b>2. How TPAI Compares.....</b>                              | <b>7</b>  |
| <b>3. Infrastructure Cost Breakdown.....</b>                  | <b>8</b>  |
| Compute (ECS Fargate).....                                    | 8         |
| Database (Aurora Serverless v2).....                          | 8         |
| <b>4. AI Model Usage.....</b>                                 | <b>9</b>  |
| Usage Model (175 Users).....                                  | 9         |
| Model Mix & Pricing (Bedrock On-Demand).....                  | 9         |
| Cost Guardrails.....                                          | 9         |
| Available Models on Bedrock.....                              | 10        |
| Cost Optimisation Levers.....                                 | 10        |
| <b>5. Networking &amp; Connectivity.....</b>                  | <b>11</b> |
| Client Account Networking (3 AZs).....                        | 11        |
| Direct Connect (Redundant).....                               | 11        |
| WAF (Web Application Firewall).....                           | 11        |
| Transit Gateway (TPAI Side).....                              | 11        |
| <b>6. Security &amp; Compliance Suite.....</b>                | <b>13</b> |
| Additional Monitoring & Infrastructure.....                   | 13        |
| <b>7. Availability &amp; Disaster Recovery.....</b>           | <b>14</b> |
| Availability Zone Strategy: 3 AZs.....                        | 14        |
| Pilot Light DR (Recommended).....                             | 14        |
| Alternative DR Strategies.....                                | 14        |
| <b>8. Reserved Capacity &amp; Savings Plans.....</b>          | <b>15</b> |
| Compute Savings Plans (Fargate).....                          | 15        |
| Total Optimisation Impact (Mid Estimate).....                 | 15        |
| <b>9. Setup &amp; Timeline.....</b>                           | <b>16</b> |
| <b>10. Cost Growth &amp; Roadmap Features.....</b>            | <b>17</b> |
| Organic Growth (Year 1 to Year 2).....                        | 17        |
| Roadmap Features & AWS Impact.....                            | 17        |
| Scaling Breakpoints.....                                      | 17        |
| <b>11. International Deployment &amp; Data Residency.....</b> | <b>18</b> |
| Regional Strategy.....                                        | 18        |
| Bedrock Model Availability by Region.....                     | 18        |
| <b>12. TPAI Managed Support.....</b>                          | <b>19</b> |
| Support Tiers.....                                            | 19        |
| What's Included.....                                          | 19        |
| What's Excluded (Billable Add-ons).....                       | 19        |

|                                                    |           |
|----------------------------------------------------|-----------|
| <b>13. Service Level Agreements.....</b>           | <b>20</b> |
| Platform Availability.....                         | 20        |
| Support Response Times.....                        | 20        |
| SLA Penalty Structure.....                         | 20        |
| <b>14. Exit Strategy &amp; Data Ownership.....</b> | <b>21</b> |
| Your Data, Your Control.....                       | 21        |
| Migration Path.....                                | 21        |
| <b>15. Investment Summary.....</b>                 | <b>22</b> |
| Monthly AWS Cost by Tier.....                      | 22        |
| Total Client Investment (Standard Support).....    | 22        |
| <b>16. Frequently Asked Questions.....</b>         | <b>22</b> |

## Executive Snapshot

*If you read only one page, read this one. For finance teams: use the companion spreadsheet to stress-test assumptions interactively.*

*Mid-range estimate: 175 users, Standard support, us-east-1, optimised. Details and assumptions follow in subsequent sections.*

**\$9,085**

AWS Monthly (optimised)

**\$30,000**

TPAI Support Monthly

**\$469K**

Total Annual

## What Drives Cost Variance

- **1. AI model mix and usage volume** (40–65% of AWS spend). Opus-heavy usage costs more; Llama/Haiku-heavy usage costs less.
- **2. AWS Support tier** (Business vs Enterprise). Enterprise commonly required by regulated customers.
- **3. Networking** (VPC endpoints + Direct Connect). Largely fixed regardless of user count.

## What You Control

- **Model governance:** Restrict Opus/Sonnet access by role or department. Route simple queries to cheaper models automatically.
- **Prompt caching:** System prompts and repeated context can be cached at 90% discount.
- **Usage budgets:** Per-user and per-department monthly token limits with alerting before thresholds are reached.

*TPAI provides a 30-day post-launch cost baseline report and quarterly cost governance reviews as part of every support contract. You will never be surprised by a bill.*

## Quick Reference: Key Terms

**Seat:** One licensed user of the TPAI platform.

**Token:** The unit AI models use to process text. Roughly 0.75 words per token. Both input (what you send) and output (what the model returns) consume tokens.

**Bedrock:** AWS's managed service for running AI foundation models. TPAI uses Bedrock so models run inside your AWS account, not on a third-party platform.

**Direct Connect (DX):** A dedicated fibre connection between your network and AWS. More reliable and secure than a VPN. Recommended for production but not required for pilots.

**Disaster Recovery (DR):** A standby environment in a second AWS region that can take over if the primary region fails. Pilot Light is the lowest-cost option; Active/Active is the highest.

## 1. Overview

*What this section covers: deployment model, key assumptions, and how to read this guide.*

*This guide provides full transparency into the AWS services, costs, and architecture decisions behind your TPAI Enterprise deployment. There are no hidden fees, no opaque bundles, and no surprises. If a number changes based on your usage, we explain exactly why and how you can control it. A companion interactive cost model spreadsheet is available so your finance team can stress-test assumptions (seats, usage, model mix, connectivity) and see totals update instantly.*

## Deployment Model

- **Single-tenant, logically air-gapped:** Deployed into your own AWS account. You retain ownership and full visibility.
- **No public internet egress:** Private subnets + VPC endpoints; SCPs prevent creation of internet/NAT gateways.
- **Secure connectivity:** Redundant Direct Connect back to TPAI's management account via Transit Gateway.
- **Your identity provider:** SAML integration with your existing IdP (Okta, Azure AD, Ping, or others).
- **Compliance-oriented:** Dedicated billing/auditing account with logging and monitoring aligned to HIPAA, SOC 2, and regulatory evidence needs.

## Key Assumptions

| Parameter          | Value                                                    |
|--------------------|----------------------------------------------------------|
| Primary region     | us-east-1 (other regions available)                      |
| DR region          | us-west-2 (Pilot Light, recommended default)             |
| Availability Zones | 3 (standard for regulated industries)                    |
| Seat count         | 100–250 (175 midpoint for estimates)                     |
| User mix           | 30% light, 50% moderate, 20% power users                 |
| AI models          | Claude (Haiku, Sonnet, Opus), Llama, Mistral via Bedrock |
| Operating hours    | Business hours (M–F) with 24/7 platform availability     |

*All estimates in this guide use the assumptions above. Your deployment may differ based on region, user count, model mix, and usage patterns. TPAI provides a customised estimate before any engagement.*



## 2. How TPAI Compares

*What this section covers: annual cost comparison against four alternatives (hire, build, buy, TPAI).*

Before reviewing detailed line items, it helps to understand how a TPAI Enterprise deployment compares to the alternatives available today. The table below compares annual cost at 175 seats.

| Alternative                                            | Annual Cost (175 seats) | Key Trade-offs                                                                        |
|--------------------------------------------------------|-------------------------|---------------------------------------------------------------------------------------|
| <b>TPAI Enterprise</b> (Standard)                      | \$469K                  | Fully managed, air-gapped, compliant, multi-model, open-source core. Deploy in weeks. |
| 2 FTEs + raw AWS                                       | \$488–588K              | 6-month ramp-up, no platform expertise, key-person risk, no compliance tooling.       |
| Enterprise AI platforms (Scale AI, Palantir AIP, etc.) | \$500K–2M+              | Per-seat licensing + infrastructure. May not be air-gapped. Long procurement cycles.  |
| Build in-house                                         | \$1–3M Year 1           | 4–6 engineers, 6–12 month build, ongoing maintenance, compliance from scratch.        |

*TPAI delivers a production-grade, compliant AI platform at a lower total cost than hiring two dedicated engineers, with no ramp-up period, no key-person risk, and a managed support contract that transfers operational risk entirely to TPAI.*

### 3. Infrastructure Cost Breakdown

*What this section covers: compute and database costs. These are relatively predictable and scale modestly with user count.*

This section details every AWS service in your deployment, how it is used, and what it costs. The largest variable is AI model usage (Section 4). Everything else is relatively predictable infrastructure.

#### Compute (ECS Fargate)

TPAI runs on AWS Fargate using Graviton (ARM64) processors for cost efficiency. The platform consists of the OWUI application layer and the Bedrock Gateway, each running as independent ECS services.

| Service             | Specification                  | Tasks | Monthly          |
|---------------------|--------------------------------|-------|------------------|
| OWUI                | 2 vCPU / 4 GB ARM64 (Graviton) | 3     | \$166            |
| Bedrock Gateway     | 1 vCPU / 2 GB ARM64 (Graviton) | 2     | \$55             |
| Fluent Bit sidecars | Bundled in task pricing        | —     | incl.            |
| <b>Subtotal</b>     |                                |       | <b>~\$221/mo</b> |

*Scales to 4–5 OWUI tasks for 250 users at peak. Business-hours scaling policies can reduce Fargate spend by 30–40% during nights and weekends.*

#### Database (Aurora Serverless v2)

Chat history and user data are stored in Aurora Serverless v2 (PostgreSQL 15.5). Chat content is encrypted at the application layer on a per-user basis via AWS KMS, providing cryptographic isolation between users on top of Aurora's standard storage encryption at rest.

| Component            | Calculation                         | Monthly          |
|----------------------|-------------------------------------|------------------|
| Writer instance      | 6 ACU avg x \$0.12/ACU-hr x 730 hrs | \$526            |
| Reader instance      | 3 ACU avg x \$0.12/ACU-hr x 730 hrs | \$263            |
| Storage              | 75 GB x \$0.10/GB                   | \$8              |
| I/O requests         | ~50M/mo x \$0.20/1M                 | \$10             |
| Backup (14-day PITR) | ~75 GB x \$0.021/GB                 | \$2              |
| <b>Subtotal</b>      |                                     | <b>~\$809/mo</b> |

*Aurora Serverless v2 scales down automatically during off-hours. Range: \$600 (100 users, light) to \$1,200 (250 users, heavy).*

## 4. AI Model Usage

*What this section covers: your largest variable cost (40–65% of AWS spend) and the controls that keep it predictable.*

*AI model usage is your largest variable cost and the one you have the most control over. Model selection, prompt design, caching strategy, and usage governance directly affect this line item. TPAI provides tools and guidance to optimise all four.*

### Usage Model (175 Users)

TPAI models three user tiers based on typical adoption patterns in regulated organisations:

| User Tier      | Users | Queries/Day | Avg Input Tokens | Avg Output Tokens |
|----------------|-------|-------------|------------------|-------------------|
| Light (30%)    | 53    | 8           | 750              | 500               |
| Moderate (50%) | 87    | 30          | 1,500            | 1,000             |
| Power (20%)    | 35    | 60          | 3,000            | 2,000             |

**Monthly totals: ~315M input / ~210M output tokens.** Estimates assume business-hours usage (~22 working days/month).

### Model Mix & Pricing (Bedrock On-Demand)

TPAI routes queries to the most cost-effective model that meets quality requirements. Your organisation controls which models are available and to whom.

| Model                      | % of Queries | Input \$/MTok | Output \$/MTok | Monthly Cost       |
|----------------------------|--------------|---------------|----------------|--------------------|
| Llama 4 Scout / Mistral 7B | 15%          | \$0.17        | \$0.66         | \$29               |
| Claude Haiku 4.5           | 20%          | \$1.00        | \$5.00         | \$273              |
| Claude Sonnet 4.6          | 40%          | \$3.00        | \$15.00        | \$1,638            |
| Claude Opus 4.6            | 25%          | \$5.00        | \$25.00        | \$1,707            |
| <b>Subtotal</b>            |              |               |                | <b>~\$3,647/mo</b> |

*Range: \$2,000 (100 light users, Llama/Haiku-heavy) to \$10,000+ (250 heavy users, Opus-heavy). Opus 4.6 on Bedrock is \$5/\$25 per MTok; direct API pricing (\$15/\$75) does not apply. Most deployments land near the midpoint once model governance and caching are enabled.*

### Cost Guardrails

TPAI provides built-in controls to ensure AI model costs stay predictable and within your approved budget:

- **Role-based model access:** Restrict premium models (Opus, Sonnet) to specific roles or departments. General users route to Haiku or Llama by default.
- **Per-user token budgets:** Set monthly token limits per user. Users receive warnings at 80% and are throttled at 100%.
- **Department budgets:** Allocate monthly token budgets by team or business unit with real-time tracking.
- **Usage alerting:** Automated alerts when usage deviates from baseline. Weekly digest reports to your designated administrator.

*If usage deviates from your approved baseline, you will know within days, not at invoice time. TPAI sets thresholds proactively and notifies on variance.*

## Available Models on Bedrock

| Family     | Model                | Input \$/MTok | Output \$/MTok | Best For                          |
|------------|----------------------|---------------|----------------|-----------------------------------|
| Meta Llama | Llama 4 Scout 17B    | \$0.17        | \$0.66         | Simple queries, summarisation     |
|            | Llama 4 Maverick 17B | \$0.24        | \$0.97         | Moderate tasks                    |
|            | Llama 3.3 70B        | \$0.72        | \$0.72         | Complex open-source alternative   |
| Mistral    | Mistral 7B           | \$0.15        | \$0.20         | Simple tasks, classification      |
|            | Mixtral 8x7B         | \$0.45        | \$0.70         | Mid-tier open-source              |
|            | Mistral Large 2      | \$8.00        | \$24.00        | Complex reasoning                 |
| Claude     | Haiku 4.5            | \$1.00        | \$5.00         | Routine tasks, high-volume        |
|            | Sonnet 4.6           | \$3.00        | \$15.00        | General-purpose, best value       |
|            | Opus 4.6             | \$5.00        | \$25.00        | Complex analysis, regulated tasks |

## Cost Optimisation Levers

TPAI implements several optimisation strategies to reduce AI model costs without affecting quality:

| Optimisation              | How It Works                                                                                 | Estimated Savings |
|---------------------------|----------------------------------------------------------------------------------------------|-------------------|
| Prompt Caching            | 90% discount on cached input token reads. Effective for system prompts and repeated context. | ~\$300–500/mo     |
| Intelligent Prompt Router | Auto-routes queries to cheaper models within the same family when quality is sufficient.     | ~\$400–600/mo     |
| Batch API                 | 50% discount for asynchronous workloads. Results returned within 24 hours.                   | Varies by volume  |
| Model governance          | Restrict premium models to specific roles/departments. Per-user token budgets.               | Highly variable   |

**Fully optimised mid-range estimate: ~\$2,647/mo** (vs. \$3,647 unoptimised). TPAI manages these optimisations as part of your support contract.

## 5. Networking & Connectivity

*What this section covers: VPC endpoints, Direct Connect, WAF, and Transit Gateway. Mostly fixed costs regardless of user count.*

The air-gapped architecture requires private connectivity for all AWS services. VPC endpoints are the largest fixed networking cost, but they are essential: they eliminate the need for NAT Gateways and ensure all traffic stays within the AWS network.

### Client Account Networking (3 AZs)

| Component                       | Detail                                    | Monthly          |
|---------------------------------|-------------------------------------------|------------------|
| VPC Interface Endpoints         | 11 endpoints x 3 AZs x \$0.01/hr x 730    | \$841            |
| ALB (OWUI)                      | Internal, 3 AZs                           | \$28             |
| ALB (Bedrock Gateway)           | Internal, 3 AZs                           | \$28             |
| Cross-AZ data transfer          | ~500 GB x \$0.01                          | \$5              |
| VPC, Subnets, Gateway Endpoints | 18 subnets (6 tiers x 3 AZs), S3/DynamoDB | \$0              |
| <b>Subtotal</b>                 |                                           | <b>~\$902/mo</b> |

### Direct Connect (Redundant)

Two redundant 1 Gbps connections provide high-availability connectivity between your network and the TPAI management account. 1 Gbps is appropriately sized for 100–250 users; LLM streaming responses are text-based and lightweight.

| Component                 | Detail              | Monthly              |
|---------------------------|---------------------|----------------------|
| DX Port (1 Gbps) x 2      | \$0.30/hr x 730 x 2 | \$438                |
| DX Data Transfer Out      | ~200 GB x \$0.02/GB | \$4                  |
| Cross-connect fees (colo) | Varies by facility  | \$200–500            |
| <b>Subtotal</b>           |                     | <b>~\$642–942/mo</b> |

### WAF (Web Application Firewall)

| Component                | Detail                                                 | Monthly         |
|--------------------------|--------------------------------------------------------|-----------------|
| Web ACL + Rules          | 1 ACL, ~10 rules (rate limiting, IP reputation, OWASP) | \$15            |
| Requests + Managed Rules | ~5M/mo + AWS Core + Known Bad Inputs                   | \$23            |
| <b>Subtotal</b>          |                                                        | <b>~\$38/mo</b> |

### Transit Gateway (TPAI Side)

The Transit Gateway on TPAI's management side connects your account to the platform's operational infrastructure. This cost is shared across all Enterprise clients and decreases per client as the portfolio grows.

| Component                    | Detail              | Monthly |
|------------------------------|---------------------|---------|
| TGW hourly + attachments (3) | 3 x \$0.05/hr x 730 | \$110   |
| Data processing              | ~200 GB x \$0.02/GB | \$4     |

| Component | Detail | Monthly   |
|-----------|--------|-----------|
| Subtotal  |        | ~\$115/mo |

## 6. Security & Compliance Suite

*What this section covers: monitoring, logging, and threat detection services. Fixed costs that satisfy audit requirements.*

A dedicated auditing account runs the monitoring, logging, and threat detection services aligned to HIPAA, SOC 2, FFIEC, and other regulatory evidence requirements. These services provide continuous monitoring, threat detection, and audit-ready logging.

| Service                | Detail                             | Monthly              |
|------------------------|------------------------------------|----------------------|
| CloudTrail (org trail) | Management + S3 data events        | \$50                 |
| AWS Config             | ~20 rules, 200 resources           | \$60                 |
| GuardDuty              | VPC Flow, DNS, CloudTrail analysis | \$150–300            |
| Security Hub           | CIS, HIPAA compliance checks       | \$30–50              |
| Macie                  | S3 sensitive data discovery        | \$50–100             |
| VPC Flow Logs          | All 18 subnets to S3               | \$30–60              |
| CloudWatch Logs        | ~50 GB ingestion/mo                | \$25                 |
| S3 (log archive)       | ~200 GB/mo, lifecycle to Glacier   | \$5–15               |
| KMS keys               | ~10 keys (per-user + service keys) | \$10                 |
| <b>Subtotal</b>        |                                    | <b>~\$410–670/mo</b> |

### Additional Monitoring & Infrastructure

| Component                                  | Monthly         |
|--------------------------------------------|-----------------|
| CloudWatch Logs (app, 30 GB)               | \$15            |
| CloudWatch Metrics + Alarms                | \$7             |
| SNS alerts, Secrets Manager, Route 53, ECR | \$12            |
| <b>Subtotal</b>                            | <b>~\$34/mo</b> |

## 7. Availability & Disaster Recovery

What this section covers: AZ strategy and DR options (Pilot Light, Warm Standby, Active/Active) with cost and recovery trade-offs.

### Availability Zone Strategy: 3 AZs

TPAI deploys across 3 Availability Zones as standard. This is the expected configuration for regulated industries and provides resilience against a single AZ failure with only 33% capacity reduction. The incremental cost over 2 AZs is approximately \$280/month (additional VPC endpoint ENIs), which is negligible relative to the compliance and resilience benefit.

### Pilot Light DR (Recommended)

For most regulated AI workloads, Pilot Light provides the optimal balance of cost and recovery capability. A minimal footprint in a secondary region maintains database replication and container images, enabling recovery within 1–4 hours.

| Component                      | Detail                        | Monthly         |
|--------------------------------|-------------------------------|-----------------|
| Aurora Global DB reader        | 1 instance, min ACU (0.5)     | \$44            |
| Aurora storage (replicated)    | ~75 GB                        | \$8             |
| ECR (replicated images)        | ~5 GB                         | \$1             |
| KMS keys + Route 53 + S3 vault | DR region infrastructure      | \$9             |
| DX Gateway attachment          | Reuse existing Direct Connect | \$37            |
| <b>Subtotal</b>                |                               | <b>~\$99/mo</b> |

RTO: 1–4 hours. RPO: minutes. Failover: deploy VPC endpoints via CDK (~15 min), promote Aurora reader (~2 min), start ECS tasks (~5–10 min), update Route 53 (~60 sec TTL).

### Alternative DR Strategies

| Strategy      | RTO       | RPO     | Monthly Cost   | When to Use                               |
|---------------|-----------|---------|----------------|-------------------------------------------|
| Pilot Light   | 1–4 hrs   | Minutes | ~\$99          | Recommended default                       |
| Warm Standby  | 15–30 min | Seconds | ~\$1,233       | If compliance mandates sub-30-min RTO     |
| Active/Active | ~0        | ~0      | ~\$7,348–8,290 | Mission-critical 24/7 with zero tolerance |

## 8. Reserved Capacity & Savings Plans

*What this section covers: how TPAI reduces your costs over time through Savings Plans, caching, and scaling optimisation.*

TPAI manages cost optimisation on your behalf as part of the support contract. The following strategies are applied based on your deployment maturity:

### Compute Savings Plans (Fargate)

| Commitment         | Discount | Monthly Fargate | Annual Savings |
|--------------------|----------|-----------------|----------------|
| On-demand          | 0%       | \$221           | —              |
| 1-year No Upfront  | ~20%     | \$177           | \$528          |
| 1-year All Upfront | ~34%     | \$146           | \$900          |
| 3-year All Upfront | ~52%     | \$106           | \$1,380        |

### Total Optimisation Impact (Mid Estimate)

| Optimisation                            | Monthly Savings    |
|-----------------------------------------|--------------------|
| Fargate Savings Plan (1-yr All Upfront) | -\$75              |
| Prompt caching (40% hit rate)           | -\$500             |
| Intelligent Prompt Router               | -\$500             |
| Off-hours ECS scaling (9–5 pattern)     | -\$60              |
| <b>Total monthly savings</b>            | <b>~\$1,135/mo</b> |

*Aurora Serverless v2 does not support Reserved Instances. Bedrock Provisioned Throughput starts at ~\$23K/month per model unit, which is not cost-effective at this usage level.*

## 9. Setup & Timeline

*What this section covers: engineering hours, deployment sequence, and the critical path to production (2–4 weeks).*

TPAI deploys your environment using infrastructure-as-code (AWS CDK). Most components are automated; the primary variable is Direct Connect provisioning and your IdP integration.

| Item                                   | Engineering Hours | Notes                                       |
|----------------------------------------|-------------------|---------------------------------------------|
| AWS account setup (Org, SCPs, billing) | 8–16              | Automated via CDK                           |
| Network stack (VPC, endpoints, NACLs)  | 4–8               | Automated via CDK                           |
| Direct Connect provisioning            | —                 | 2–6 week lead time (facility-dependent)     |
| Aurora database deployment             | 4–8               | CDK + initial schema                        |
| ECS service deployment                 | 2–4               | Image pull via Transit Gateway              |
| Compliance suite configuration         | 8–16              | CloudTrail, Config, GuardDuty, Security Hub |
| SAML integration                       | 4–16              | Depends on IdP (Okta, Azure AD, Ping)       |
| WAF rule tuning                        | 4–8               | Baseline + client-specific rules            |
| DR setup (Pilot Light)                 | 4–8               | Aurora Global DB, ECR replication           |
| Acceptance testing & handoff           | 8–16              | End-to-end validation                       |
| <b>Total</b>                           | <b>50–100 hrs</b> | <b>~2–4 weeks elapsed</b>                   |

*Direct Connect provisioning (LOA process, 2–6 weeks) is the longest lead item. All other work proceeds in parallel.*

## 10. Cost Growth & Roadmap Features

*What this section covers: how costs evolve as usage matures and new platform features (web search, deep research, API integrations) are adopted.*

Your AWS costs will grow as your team adopts AI more deeply. This is healthy: it means the platform is delivering value. TPAI monitors usage trends and provides quarterly cost reviews as part of your support contract.

### Organic Growth (Year 1 to Year 2)

| Factor                                  | Impact                                         |
|-----------------------------------------|------------------------------------------------|
| User count growth (+50–100 users)       | +\$2,000–4,000/mo Bedrock                      |
| Adoption maturity (queries/day +20–30%) | +\$700–1,100/mo Bedrock                        |
| Document/RAG storage growth             | +\$5–12/mo S3                                  |
| Offset: optimisation maturity           | -\$500–1,000/mo (caching, routing, governance) |

### Roadmap Features & AWS Impact

TPAI's product roadmap includes features that will increase the value your team extracts from AI. Each feature increases Bedrock usage proportionally:

| Feature                                  | Low Estimate       | High Estimate       | Timeline   |
|------------------------------------------|--------------------|---------------------|------------|
| Web Search                               | +\$265/mo          | +\$1,040/mo         | Q2 2026    |
| Deep Research                            | +\$1,510/mo        | +\$7,030/mo         | Q3 2026    |
| API Integrations (email, docs, calendar) | +\$530/mo          | +\$3,095/mo         | Q3–Q4 2026 |
| Automation & Agent Workflows             | +\$515/mo          | +\$4,050/mo         | Q4 2026    |
| <b>Total roadmap impact</b>              | <b>+\$2,820/mo</b> | <b>+\$15,215/mo</b> |            |

**Year 2 projection with features: ~\$10,000–22,000/mo** (vs. ~\$7,100 base Year 1). These features drive significant value for your organisation and proportionally increase Bedrock spend.

### Scaling Breakpoints

| Milestone   | Action Required                   | Cost Impact         |
|-------------|-----------------------------------|---------------------|
| 250+ users  | Add 4th–5th OWUI task             | +\$55–110/mo        |
| 300+ users  | Increase Aurora ACU floor         | +\$100–200/mo       |
| 500+ users  | Evaluate provisioned Aurora + RIs | Potential -\$200/mo |
| 1000+ users | Evaluate active/active DR         | +\$7,000+/mo        |

## 11. International Deployment & Data Residency

*What this section covers: regional AWS options, Bedrock model availability by geography, and cost variations for non-US deployments.*

TPAI deploys into AWS regions with Bedrock support. The architecture is region-agnostic; the same CDK stacks deploy to any supported region. The deployment supports data residency objectives by keeping all processing within the selected region; jurisdictional obligations remain subject to customer contract and applicable law.

### Regional Strategy

| Region         | AWS Region                 | Primary Use Case              |
|----------------|----------------------------|-------------------------------|
| North America  | us-east-1                  | SaaS + Enterprise (US/Canada) |
| Europe         | eu-west-1 / eu-central-1   | Enterprise (EU)               |
| Middle East    | me-south-1 / me-central-1  | Enterprise (GCC)              |
| Africa         | af-south-1 (Cape Town)     | Enterprise                    |
| Southeast Asia | ap-southeast-1 (Singapore) | Enterprise                    |

### Bedrock Model Availability by Region

| Model             | US        | EU      | Middle East   | Africa        | APAC          |
|-------------------|-----------|---------|---------------|---------------|---------------|
| Claude Opus 4.6   | In-region | EU CRIS | In-region     | In-region     | In-region     |
| Claude Sonnet 4.6 | In-region | EU CRIS | In-region     | In-region     | In-region     |
| Claude Haiku 4.5  | In-region | EU CRIS | In-region     | In-region     | In-region     |
| Llama / Mistral   | In-region | Limited | Not available | Not available | Not available |

*Middle East, Africa, and APAC deployments may be Claude-only, increasing model costs ~15–20% vs. US baseline. EU deployments typically carry a 5–10% infrastructure premium. Direct Connect costs vary significantly by region.*

## 12. TPAI Managed Support

*What this section covers: support tiers, what is included, what is excluded, and pricing for deployments above 250 seats.*

TPAI provides fully managed operations for your deployment. Your team uses the platform; our team runs it. This transfers operational complexity and risk to TPAI while ensuring your deployment stays secure, compliant, and optimised.

### Support Tiers

|                | Standard (\$30,000/mo) | Premium (\$40,000/mo) |
|----------------|------------------------|-----------------------|
| Support hours  | Business hours (M–F)   | Extended hours        |
| PI response    | 2 hours                | 1 hour                |
| Patching       | Monthly                | Weekly                |
| Reviews        | Quarterly              | Monthly               |
| DR drills      | Annual                 | Quarterly             |
| Named engineer | No                     | Yes                   |

*For deployments above 250 seats, per-seat pricing: ~\$120/seat/mo (Standard) or ~\$160/seat/mo (Premium).*

### What's Included

- Infrastructure monitoring and alerting (24/7 automated; human response per tier)
- Patch management and security updates (OS, containers, dependencies)
- Software upgrades (OWUI, Bedrock Gateway, CDK stacks)
- Cost optimisation and model usage analysis
- Compliance support (audit preparation, evidence collection, control documentation)
- DR testing and failover drills (frequency per tier)
- Capacity planning and scaling recommendations
- SCP management and policy updates
- AWS Support case management

### What's Excluded (Billable Add-ons)

| Service                                             | Rate                   |
|-----------------------------------------------------|------------------------|
| Custom feature development                          | \$250–350/hr           |
| Architecture changes beyond platform                | \$250–350/hr           |
| Compliance certification processes (SOC 2, HITRUST) | Scoped per engagement  |
| Onboarding and training                             | \$5,000–10,000 package |
| Bespoke AI integration engineering                  | Scoped per engagement  |

## 13. Service Level Agreements

*What this section covers: uptime targets, response times, penalty structure, and what regulated clients should expect.*

### Platform Availability

| SLA Tier | Uptime Target | Measured During          | Annual Downtime Allowed |
|----------|---------------|--------------------------|-------------------------|
| Standard | 99.5%         | Business hours (M–F 9–5) | ~13 hours/year          |
| Premium  | 99.9%         | Business hours           | ~2.6 hours/year         |

*The platform runs 24/7 but the SLA is measured during defined business hours. Scheduled maintenance outside business hours does not count against the SLA.*

### Support Response Times

| Priority      | Definition                 | Standard          | Premium         |
|---------------|----------------------------|-------------------|-----------------|
| P1 (Critical) | Platform fully unavailable | 2-hour response   | 1-hour response |
| P2 (Major)    | Significant degradation    | 4-hour response   | 2-hour response |
| P3 (Minor)    | Non-critical issue         | Next business day | 4-hour response |
| P4 (Request)  | Enhancement / question     | 2 business days   | 1 business day  |

### SLA Penalty Structure

| Uptime Achieved    | Service Credit                   |
|--------------------|----------------------------------|
| 99.0% – SLA target | 5% of monthly support fee        |
| 95.0% – 99.0%      | 10% of monthly support fee       |
| 90.0% – 95.0%      | 20% of monthly support fee       |
| Below 90.0%        | 25% of monthly support fee (cap) |

*Credits capped at 25% of monthly support fees. Exclusions: force majeure, scheduled maintenance (72-hr notice), client-caused outages, AWS region-wide outages. Three SLA breaches in 6 months triggers client right to terminate with 30-day notice.*

## 14. Exit Strategy & Data Ownership

*What this section covers: your data rights, migration path, and contractual protections. No proprietary lock-in.*

*You own your data. You own your AWS account. If you leave TPAI, everything stays with you. There is no proprietary lock-in.*

### Your Data, Your Control

| Data                         | Location                           | Portability             |
|------------------------------|------------------------------------|-------------------------|
| Chat history & conversations | Aurora PostgreSQL (OWUI database)  | Standard pg_dump export |
| User accounts & preferences  | Aurora PostgreSQL                  | Standard pg_dump export |
| Uploaded documents           | S3 bucket in your account          | Direct S3 access / copy |
| Audit logs                   | S3 + CloudWatch in billing account | Standard export         |
| Configuration & settings     | Aurora + Secrets Manager           | Exportable              |

### Migration Path

- **Full data export:** PostgreSQL dump + S3 contents. TPAI provides export tooling and documentation.
- **Open-source core:** Open WebUI is open-source. You can self-host or migrate to another provider. No proprietary formats or APIs.
- **Account transfer:** Your AWS account can be migrated to your own Organisation. All resources remain intact.
- **Transition period:** 30–90 day transition where TPAI continues to operate while you establish an alternative.
- **Data deletion:** TPAI deletes all client data within 30 days of transition completion.

## 15. Investment Summary

What this section covers: consolidated monthly and annual totals across three deployment sizes.

### Monthly AWS Cost by Tier

| Category                                  | Low (100 users) | Mid (175 users) | High (250 users) |
|-------------------------------------------|-----------------|-----------------|------------------|
| Bedrock API (LLM)                         | \$2,000         | \$3,647         | \$10,000         |
| Networking                                | \$880           | \$902           | \$930            |
| Aurora Serverless v2                      | \$600           | \$809           | \$1,200          |
| Direct Connect (redundant)                | \$642           | \$792           | \$942            |
| Compliance Suite                          | \$350           | \$540           | \$670            |
| ECS Fargate                               | \$150           | \$221           | \$350            |
| AWS Support                               | \$2,000         | \$3,000         | \$4,000          |
| Other (TGW, DR, WAF, monitoring)          | \$296           | \$294           | \$312            |
| <b>AWS Total</b>                          | <b>\$6,918</b>  | <b>\$10,220</b> | <b>\$18,426</b>  |
| <b>Optimised (caching + SP + scaling)</b> | <b>\$5,843</b>  | <b>\$9,085</b>  | <b>\$16,791</b>  |

### Total Client Investment (Standard Support)

| Component               | Low (100)       | Mid (175)       | High (250)      |
|-------------------------|-----------------|-----------------|-----------------|
| AWS (optimised)         | \$5,843         | \$9,085         | \$16,791        |
| TPAI Support (Standard) | \$30,000        | \$30,000        | \$30,000        |
| <b>Total Monthly</b>    | <b>\$35,843</b> | <b>\$39,085</b> | <b>\$46,791</b> |
| <b>Per Seat/Month</b>   | <b>\$358</b>    | <b>\$223</b>    | <b>\$187</b>    |
| <b>Annual</b>           | <b>\$430K</b>   | <b>\$469K</b>   | <b>\$561K</b>   |

At 175 seats, TPAI Enterprise costs less than hiring two specialised engineers, deploys in weeks instead of months, transfers operational risk entirely to TPAI, and provides a compliance-ready platform from day one.

## 16. Frequently Asked Questions

What this section covers: the six questions procurement and finance teams most commonly ask.

### What makes the bill go up?

AI model usage (Bedrock) is the primary variable. More users, heavier usage, or a shift toward premium models (Opus) increases token consumption. User growth, adoption of new features (web search, deep research, API integrations), and expanded document processing all contribute. Infrastructure costs (networking, compute, database) grow modestly and predictably with user count.

## What can we cap?

Per-user token budgets, department budgets, and role-based model access all provide hard limits. You can restrict Opus to specific roles, set monthly token ceilings per user, and receive alerts before thresholds are reached. TPAI configures these as part of onboarding.

## What is optional vs recommended?

- **Optional:** Direct Connect (can start with VPN for POC), Premium support tier, Warm Standby/Active-Active DR, Macie (S3 data discovery).
- **Recommended:** 3 AZs, Pilot Light DR, Enterprise AWS Support, redundant Direct Connect for production, full compliance suite.

## What changes between 100, 175, and 250 seats?

Bedrock costs scale roughly linearly with users. Infrastructure (networking, compute, database) increases modestly. The biggest delta is Bedrock: \$2,000/mo at 100 light users vs \$10,000+/mo at 250 heavy users. Per-seat cost decreases as you add users because fixed infrastructure is spread across more seats (\$358/seat at 100 vs \$187/seat at 250).

## What does TPAI Support cover vs AWS Support?

TPAI Support covers everything about your deployment: monitoring, patching, upgrades, compliance, DR testing, cost optimisation, and capacity planning. AWS Support covers the underlying AWS services (infrastructure incidents, service limits, architecture reviews). TPAI manages the AWS Support relationship on your behalf, including case escalation and TAM coordination.

## What happens when new features launch?

Features like web search, deep research, and API integrations increase Bedrock usage proportionally to adoption. TPAI provides a cost impact estimate before enabling any new feature and includes the projection in your quarterly cost review. No feature is enabled without your approval.

*Your Data. Your AI. Totally Private.*

© 2026 Totally Private AI, Inc. All rights reserved.