

Private AI Service

Enterprise AI, On Your Terms

Your organization's dedicated, on-premises AI infrastructure — enterprise-grade intelligence, complete data sovereignty, and unlimited usage for up to **60 users**, all for one predictable annual investment.

GSCX Group's **Private AI Service** delivers a fully managed, on-premises large language model environment with dedicated enterprise-grade AI compute — 1 petaFLOP of AI performance, 128 GB of unified memory, and support for models up to 200 billion parameters. No rate limits, no token caps, no data sent to third-party clouds. Purpose-built AI for productivity, coding, research, and operations — under a single flat-rate annual plan.

1PF

AI COMPUTE

200B

MAX MODEL PARAMS

60

USERS / UNIT

6+

BUNDLED AI APPS



100% PRIVATE & ON-PREMISES

All AI inference runs within your environment. Data never leaves your facility — no third-party cloud, no API calls, no training on your data.



UNLIMITED USAGE

No per-token billing, no query caps, no overage surprises. Run as many prompts as your team needs — analytics, drafts, code, and more.



UP TO 60 NAMED USERS

Each unit supports 60 concurrent seats with full model access — general purpose AI and coding AI included under a single annual license.

Included AI Models



General Purpose AI Model

Handles writing, research, summarization, data analysis, customer communications, document drafting, and complex multi-step reasoning across your entire organization.

Reasoning Writing Analysis Q&A
Summarization



Coding & Development Model

Accelerates software teams with AI pair-programming — code generation, automated debugging, code review, documentation, and architecture guidance without exposing source code externally.

Code Gen Debugging Review Docs
Architecture



Ready to Get Started?

Contact our team to discuss your organization's requirements, user count, and deployment timeline. Custom quotes available for multi-unit deployments.



Service Infrastructure — Dedicated AI Compute Specifications

Every Private AI Service License is backed by **dedicated, enterprise-grade AI compute** provisioned exclusively for your organization. Your environment delivers 1 petaFLOP of AI performance, 128 GB of unified memory, and full support for models up to 200B parameters — entirely on-premises, with zero shared resources and zero data leaving your network.

Dedicated AI Compute Specifications



AI PERFORMANCE

1 petaFLOP INT8

Enterprise-grade AI compute purpose-built for model inference and training at scale — delivering performance previously accessible only through expensive cloud GPU rentals.



PROCESSING

20-Core ARM

20 high-performance cores handling complex AI preprocessing, multi-user orchestration, and concurrent workloads — all within your dedicated environment.



UNIFIED MEMORY

128 GB

High-bandwidth unified memory shared across all AI compute enables large-scale model prototyping and fine-tuning that would otherwise require a multi-server setup.



STORAGE

4 TB NVMe

Ultra-fast NVMe solid-state storage ensures rapid dataset access and model loading. Your entire AI data estate stays local — no cold-start latency from cloud storage.



CONNECTIVITY

24/7/365

Always-on service availability with continuous uptime — your AI environment is accessible around the clock, every day of the year, with no scheduled downtime windows.



ISOLATION

100% Dedicated

Your compute environment is never shared. All resources — memory, storage, processing — are reserved exclusively for your organization's workloads around the clock.



Supports Models Up to 200 Billion Parameters — Entirely On-Premises

The service's pre-configured AI stack supports local deployment of LLMs up to 200B parameters. Run Llama 3, Mistral, Phi-3, or your own fine-tuned models — without a single API call leaving your network. Your data never touches a public cloud.



Production-Grade AI Performance

Your dedicated compute delivers the same class of AI performance powering the world's most advanced data-center AI clusters — enterprise inference and training at massive scale, without the cloud price tag.



Energy Efficient & Low Operational Cost

The unified memory architecture dramatically reduces power overhead vs. discrete multi-server setups. Run 24/7 AI workloads without data-center-grade power or cooling infrastructure.



Dedicated Enterprise AI Compute at a Fraction of the Market Cost

Comparable dedicated AI infrastructure sourced independently starts at **\$149,000 or more for compute alone**

, plus software, licensing, and IT overhead. Dedicated GPU rack co-location runs \$120,000–\$240,000 per year in recurring fees. The GSCX Group Private AI Service delivers equivalent mission-critical AI performance — compute, software, and support together — at a fraction of those costs.



Enterprise Certified

Compute stack certified for enterprise workloads



Air-Gap Ready

Fully operational in isolated networks. No



Pre-Configured

AI compute stack arrives fully configured. First



Horizontally Scalable

Add service units to expand capacity. Each

with production-grade
drivers and AI runtime.

internet dependency for
AI inference.

model running in hours,
not weeks.

unit adds 60 seats and 1
petaFLOP of compute.

Bundled Software Suite — Six Enterprise AI Applications Included

Every Private AI Service License includes six production-ready enterprise AI applications that run **entirely within your dedicated environment** — unlimited usage, zero per-seat licensing fees, zero data sent externally. These applications would cost **\$50,000–\$150,000+ annually** if licensed separately from competing vendors. They are all included in your annual License.

REASONING ENGINE

Brainiac

Enterprise AI reasoning and decision-support platform. Connects to your internal data sources, performs multi-step analysis, and generates executive-ready intelligence — without data ever leaving your premises.

WEB INTELLIGENCE

Website Agent

Autonomous AI agent that monitors, analyzes, and engages with web properties. Ideal for competitive intelligence, customer-facing chatbot deployment, and automated digital research workflows.

KNOWLEDGE & TRAINING

Quizio

AI-powered knowledge assessment and adaptive training platform. Build learning programs, automate employee upskilling, and measure knowledge retention across your organization at scale.

COMMUNICATIONS AI

Email Agent

Intelligent email automation that reads context, drafts responses, triages inbound communications, and manages follow-ups — reducing email overhead by 60–80% across your enterprise without external data exposure.

GENERATIVE AI CHAT

LLM Chat

A private, enterprise-grade chat interface powered by your locally hosted large language models. Equivalent to ChatGPT Enterprise — all conversations are fully private, auditable, and never leave your network.

DEVELOPER TOOLS

Coding Assistant

AI pair-programmer that understands your private codebase. Automates code review, generates boilerplate, identifies bugs, and accelerates developer velocity — without sending source code to external APIs.



Marketplace.jozieai.com — Expand Your Capabilities

Access an ever-growing ecosystem of additional AI applications, domain-specific models, workflow integrations, and industry-specific LLMs through the GSCX Group Marketplace. From vertical-specific AI agents to custom automation workflows — your service grows with your business needs.

Software Value vs. Standalone Licensing



LLM CHAT (CHATGPT ENTERPRISE EQUIV.)

ChatGPT Enterprise costs **\$30–60/user/month**. For 60 users that is **\$21,600–\$43,200/year** — for a single tool, with your data handled by OpenAI.



CODING ASSISTANT (GITHUB COPILOT EQUIV.)

GitHub Copilot Business costs **\$19/user/month**. For a 60-person engineering team that is **\$13,680/year** — for one tool, with source code sent to Microsoft servers.



EMAIL & WEB AGENTS (AUTOMATION EQUIV.)

Enterprise email AI and web automation platforms (e.g., Salesforce Einstein, HubSpot AI) typically run **\$25,000–\$80,000/year** for comparable capability.

◆ All Six Applications. One Price. Zero External Data Exposure.

The combined market value of the six bundled applications — licensed separately from enterprise vendors — exceeds **\$80,000–\$180,000 per year**

for a 60-user organization. With Private AI Service, you receive all six plus dedicated compute infrastructure under a single annual License, with every query staying fully within your controlled environment.

Return on Investment — The Math Is Undeniable

Cloud AI compute costs are severe at enterprise scale. GSCX Group's **Private AI Service** provides the same capability as infrastructure that comparable enterprises currently pay **\$500,000 to \$2,000,000+ annually** to access via cloud providers — at a dramatically lower annual cost. The first-year savings alone are transformational.

CLOUD AI — ANNUAL COST

\$1,200,000

Estimated average enterprise spend on cloud GPU compute (AWS/Azure/GCP) for equivalent AI performance

\$83K

Per Month

\$2,740

Per Day

∞

Recurring Forever

PRIVATE AI SERVICE — ANNUAL COST

Contact Us

Flat annual fee — compute, software, infrastructure, and support all included. No usage overages.

\$0

Usage Overages

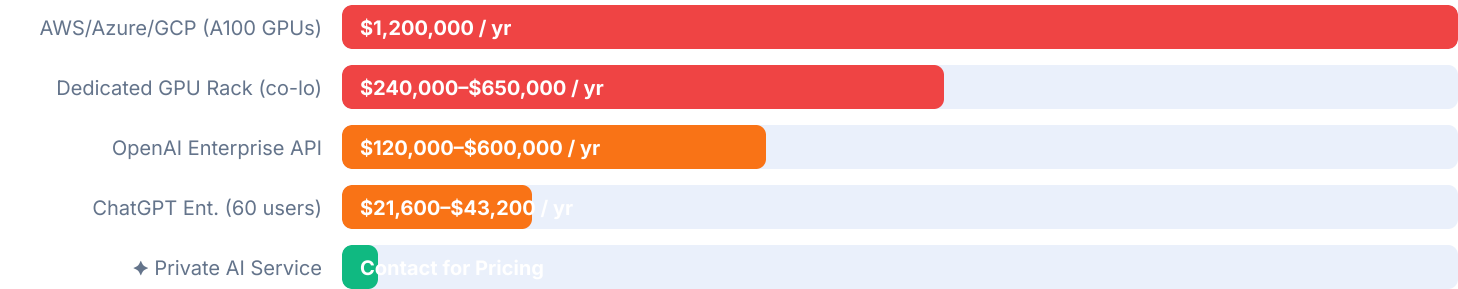
60

Users Included

6+

Bundled Apps

Annual Cost Comparison by Deployment Type



SOLUTION	ANNUAL COST	3-YEAR TOTAL	INFRASTRUCTURE	DATA PRIVACY	USERS
Cloud GPU (AWS/Azure/GCP)	\$500K–\$2M+	\$1.5M–\$6M+	Shared/Rented	Vendor-Held	Metered
OpenAI / Anthropic API	\$120K–\$600K	\$360K–\$1.8M	None	Third-Party	Per Token
ChatGPT Enterprise (60 users)	\$21,600–\$43,200	\$65K–\$130K	None	OpenAI Servers	60 seats
Dedicated Co-Lo GPU Rack	\$120K–\$240K	\$440K–\$920K	Co-located	Co-Lo Risk	Varies
◆ GSCX Group Private AI Service	Contact Us	Contact Us	Included	100% On-Premises	60 / unit

AWS P4D.24XLARGE (8× A100)

\$32/hr → \$280K/yr

Equivalent cloud GPU capacity billed hourly. At 24/7 enterprise utilization this exceeds \$280,000 annually — every year, compounding.

Perpetual Recurring

OPENAI ENTERPRISE LICENSE

\$30–60/user/mo

For 60 users: \$21,600–\$43,200/yr. No model control, no fine-tuning rights, no data privacy. Prices historically increase annually.

Vendor Lock-In

◆ PRIVATE AI SERVICE

Contact Us

1 petaFLOP compute, 128GB unified memory, 200B parameter support, 6 enterprise apps, and 60 user seats — all on your premises.

Best Total Value

Who Benefits & How to Get Started

Enterprise Use Cases



Private LLM Deployment

Run Llama 3, Mistral, Phi-3, or custom fine-tuned models locally. Serve 60 internal users with zero per-query cost and full data sovereignty — ideal for legal, finance, and healthcare.



AI Research & Prototyping

128 GB of unified memory and 1 petaFLOP of dedicated AI compute lets your data science team fine-tune and prototype models that previously required \$50K/month in cloud GPU time.



Business Intelligence & Analytics

Connect Brainiac to your ERP, BI tools, and databases. Generate executive analysis and surface hidden patterns — powered by local AI, with no data leakage.



Internal AI Assistant Rollout

Deploy LLM Chat and Coding Assistant across your workforce. Give every employee a private ChatGPT-equivalent — with your proprietary data embedded — for one flat annual fee.



Communications Automation

Email Agent and Website Agent automate high-volume repetitive communication that consumes hours daily. Triage, draft, respond, and escalate — without adding headcount.



Workforce Training & Upskilling

Quizio deploys AI-adaptive training programs across your organization. Continuously assess knowledge gaps, personalize learning paths, and certify employees — all running privately within your service environment.

Ideal Buyer Profiles



FINANCIAL SERVICES

Banks, Fintechs & Asset Managers

Data residency regulations make cloud AI non-starters. Private AI Service delivers institutional-grade AI within your compliance perimeter.

- Fraud detection inference
- Private document analysis
- Regulatory automation



HEALTHCARE & LIFE SCIENCES

Hospitals, Clinics & Pharma

HIPAA and GDPR demand patient data never leaves controlled environments. Local AI makes clinical intelligence viable without regulatory risk.

- Clinical note summarization
- Drug interaction analysis
- Medical imaging prep



LEGAL & PROFESSIONAL SERVICES

Law Firms & Consulting Groups

Client confidentiality is non-negotiable. Local AI keeps privileged documents privileged while unlocking AI-assisted legal research and drafting.

- Contract review & redlining
- Case research automation
- Document summarization



MANUFACTURING & ENGINEERING

Industrial & Defense Contractors

Proprietary designs cannot be shared with cloud vendors. The on-premises deployment model enables edge AI on the factory floor or in field deployments.

- Predictive maintenance
- Quality control AI
- Supply chain optimization



EDUCATION & RESEARCH

Universities & Research Labs

Academic budgets make cloud GPU costs prohibitive. A single unit gives an entire department dedicated petaFLOP-class AI for less than one month of cloud spend.

- NLP and ML research
- Student AI lab access
- Institutional AI tools



TECHNOLOGY & SAAS

AI-Native Startups & Scale-Ups

Stop burning runway on cloud GPU bills. Use Private AI Service as your development and staging environment — iterate privately, scale to cloud only for production peaks.

- LLM product development
- Model fine-tuning pipelines
- Offline-first AI features

Private AI Service Plan

Contact for Pricing

Flat-rate · No per-token overages · No per-seat add-ons

- ✓ Dedicated on-premises AI compute environment
- ✓ 1 petaFLOP INT8 AI performance
- ✓ 128 GB unified memory + 4 TB NVMe storage
- ✓ Up to 200B parameter model support
- ✓ Up to 60 named user seats
- ✓ Unlimited token usage — no caps
- ✓ General Purpose & Coding AI Models
- ✓ 6 bundled enterprise AI applications
- ✓ GSCX Group Marketplace access
- ✓ Onboarding & ongoing support