Gyana MCP Servers

INTRODUCTION

The Hidden Cost of Building AI in Finance

The problem: Financial institutions face three recurring AI-infrastructure challenges:

- **Vendor Lock-In** — Dependence on a single LLM provider.
- **Integration Chaos** — Voice, data, and AI layers don't communicate.
- **Compliance Risk** — Black-box AI fails regulatory tests.

Every fintech AI project begins with one question: which LLM do we bet our infrastructure on?

- Choose wrong — you're locked in.
- Build everything yourself — you're 6 months behind.

Gyana MCP Server eliminates both problems.

Introducing Gyana MCP Server

🤖 Universal AI – Orchestrates multiple models (GPT, Claude, Gemini, Llama) through a single MCP-compliant API.

🎙️ Universal Voice – Complete STT → LLM → TTS pipeline enabling natural, voice-native interaction.

🧠 Universal Vector Knowledge Base – Persistent, semantic memory that brings organisational context to every task.

One Platform. Three Universal Components. Zero Vendor Lock-In.

Think of it as AI-Infrastructure-as-a-Service. Not a chatbot. Not another wrapper.

An AI infrastructure layer designed for interoperability, compliance, and scale — the foundation to build AI-enabled applications.

Gyana Universal AI MCP Server

A unified WebSocket-based MCP (Model Context Protocol) server that enables access to multiple AI providers (OpenAI, Anthropic, Gemini) through a single endpoint with secure access, usage tracking, and standardised request/response formatting.

✨ Key Benefits

✅ Single endpoint for multiple AI providers

✅ Built-in usage tracking and rate limiting

✅ Standardised request/response format across all providers

✅ Authentication and user management

✅ No need to manage multiple API keys - just one access key for all providers

✅ WebSocket-based for real-time communication

✅ MCP protocol compliance for seamless integration with MCP clients

Provider-agnostic - switch between providers without code changes

Gyana Universal Voice MCP Server

A unified WebSocket-based MCP (Model Context Protocol) server for end-to-end voice processing: Speech-to-Text → Enhance Prompt with RAG + AI Processing (with child safety) → Text-to-Speech through a single endpoint with secure access, usage tracking, and multi-provider support.

✨ Key Benefits

✅ Complete voice pipeline in one API call

✅ Built-in child safety at every stage

✅ Multi-provider support with dynamic routing

✅ Conversation continuity

✅ Usage tracking and tier management

✅ No need to manage multiple API keys for different providers

✅ WebSocket-based for real-time communication

✅ MCP protocol compliance for seamless integration

✅ Base64 encoding/decoding handled automatically

✅ Support for WAV, OGG, and MP3 formats

✅ Customizable system prompts

✅ Enhance prompts with RAG Implementation

Provider-agnostic - switch between providers without code changes

Gyana Universal Vector KB MCP Server

A unified WebSocket-based MCP (Model Context Protocol) server for building and searching vector knowledge bases from URLs through a single endpoint with secure access, usage tracking, and automatic vector database export.

✨ Key Benefis

- ✅ Complete RAG pipeline in one API call
- ✅ Automatic URL fetching and content extraction
- ✅ Built-in chunking and embedding generation
- ✅ Vector database export for local use
- ✅ Usage tracking and tier management
- ✅ WebSocket-based for real-time communication
- ✅ MCP protocol compliance for seamless integration
- ✅ Base64 encoding/decoding handled automatically
- ✅ ChromaDB format
- ✅ Semantic search with relevance scoring

Provider-agnostic - switch between providers without code changes

Task-Driven Today, Agentic-Ready Tomorrow

Today

- Gyana MCP Server runs in a task-driven mode — human-supervised, auditable, and compliant by design.
- Every request is enriched by the Vector Knowledge Base for context, and it operates within clear governance boundaries for traceability.

Future-Ready

- The same infrastructure that powers task-driven AI today can evolve into agentic AI tomorrow.
- Start with controlled, auditable task automation, then add orchestration and sequencing as confidence grows.
- Introduce semi-autonomous agents only when governance and compliance are in place.

From controlled AI-enabled automation to agentic AI — on your terms!

Real Value | Time & Cost Savings

ROI Example:
Traditional build = 5–8 months

vs.

Gyana = 1-2 weeks / $0 in infrastructure build.

Without Gyana

- Multi-LLM integration (1–2months)
- e2e Voice pipeline (3–4 months)
- Create / Search Vector KB (1–2 months)
- Failover logic & cost tracking (ongoing)

With Gyana

- Production-ready Day 1
- Add voice via 1 API call
- Deploy in minutes
- Automated orchestration

Stop building infrastructure. Start building products.

Our Philosophy Is Different

The AI ecosystem (and many others related) changes every six months; architecture must last for years.

That’s why we’ve designed askAITHENA as an **intelligence infrastructure**, not a tool.

- The future belongs to systems that are **architecture-first**, **agnostic**, and **composable**.
- That means our platform isn’t a fancy front end — it’s the connective tissue between AI models, data layers, APIs, and decision systems.
- Our value lies in the **design, orchestration, and adaptability** we bring — not in markup over model usage.
- Our subscriptions reflect our **added intelligence** — the architecture, logic engines, and layered evaluation frameworks we’ve built.
- Model usage itself remains **pass-through or top-up**, transparently linked to provider cost.

askAITHENA — where architecture-first design and applied AI converge to engineer intelligent systems that optimise performance today while ensuring resilience for the future.

CONTACT

DM on WA+ : +65 91686869

Link to askAITHENA Email

Link to askAITHENA website