qgpt.qbadvisory.com

Enterprise Knowledge – Powered by AI

A sovereign air-gapped Enterprise LLM platform that serves as a full AI-powered assistant integrating with almost all sources of Enterprise Data plus a multi-modal inference engine used per purpose and usage patterns

Smart • Secure • Sovereign

Executive Summary

Q-GPT is an Enterprise AI knowledge platform enabling secure conversational AI across the entire enterprise ecosystem. Built on open-source technologies advanced Retrieval Augmented Generation **(RAG)** -delivering accurate, cited responses with full data sovereignty.

Q-GPT powers business-critical intelligence, governance, and automation. It is designed to work with multiple models and internal vector database. Its innovation combines hybrid search, contextual retrieval, **vLLM** inference, and a forward roadmap featuring **StructRAG (Q2 2026)** and **RLM (Q3 2026)**.

40+ Data Sources

Connects to SharePoint, SAP S4/HANA, Microsoft D365, Confluence, Salesforce, Slack, Outlook, G-Suite and more.

Hosting – On-prem/ Hybrid/ Private Cloud

Enterprise data never leaves your environment. Complete sovereignty and regulated.

EU Compliant

Built for EU AI Act, DORA and GDPR regulations.

Open Source

No vendor lock-in. MIT Licensed core platform. No separate subscription costs other than token fees for chosen models.

White Labeling

Q-GPT can be white labeled by enterprises with their own branding and appearance

https://qgpt.qbadvisory.com

The Problem(s)

- ❌ Unassembled Enterprise Knowledge
- ❌ Tribal dependency
- ❌ Inefficient audit & compliance
- ❌ Uncontrolled AI usage risk
- ❌ Frenzied investments on AI initiatives
- ❌ Risk concentration with vendor lock-in
- ❌ Data sovereignty not respected

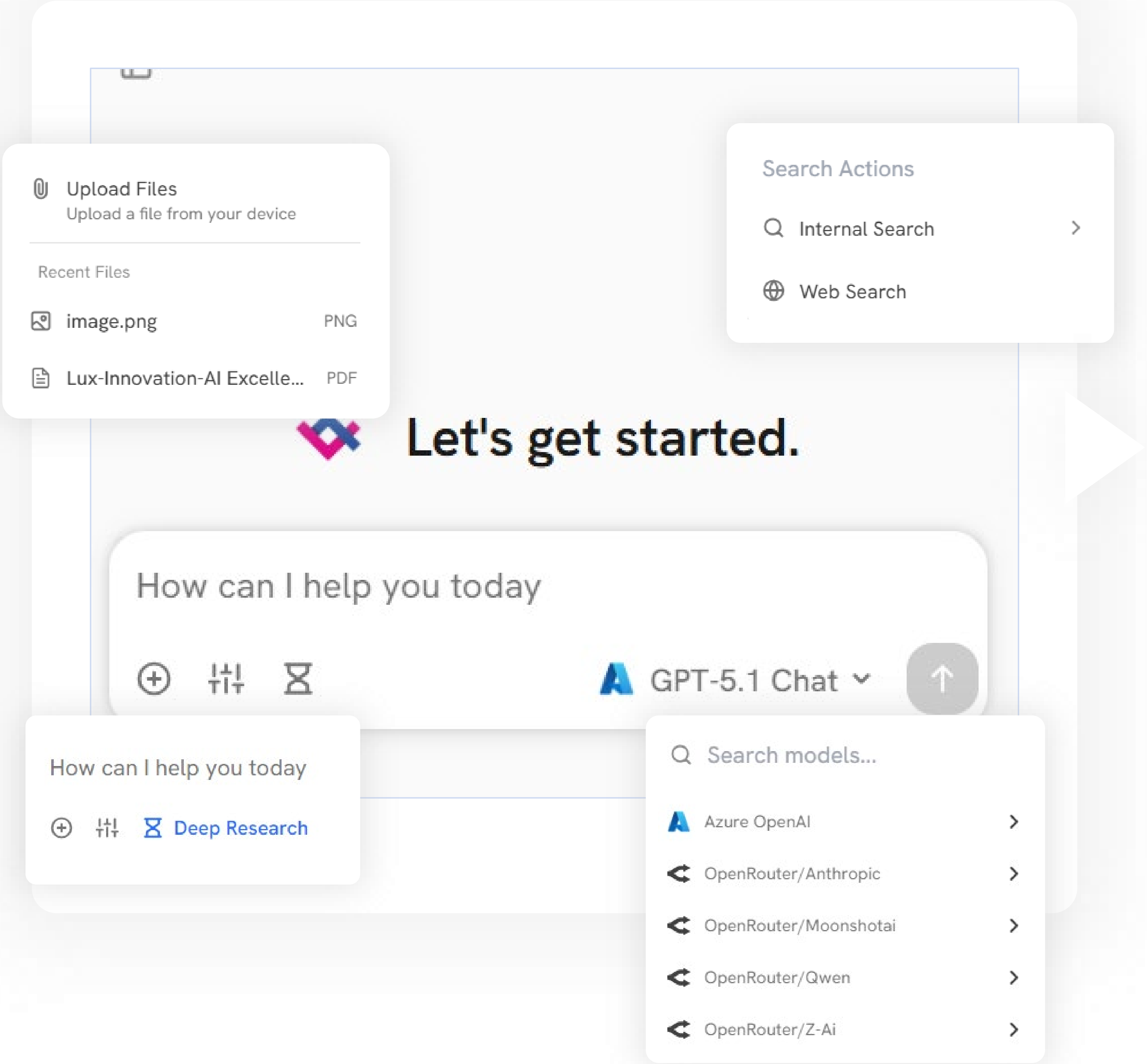
Solutions

- Unified Enterprise Intelligence
- Governed AI with source level ACLs
- Indexed, retrievable, reasoning-ready knowledge
- Reduced dependency on tribal knowledge
- Controlled innovation without regulatory risk
- Air-gapped LLM access to preserve data sovereignty
- Inhouse RAG and Prompt Orchestration
- Using models for purpose reducing hallucinations
- De-coupled Architecture allows independent scaling
- Controlled TCO for AI investments

https://qgpt.qbadvisory.com

qba

Q-GPT Introduction



Unified Knowledge Access

Native connectors to **40+ enterprise systems** *(Microsoft 365, Google Workspace, Atlassian, Salesforce, SAP, D365, databases, and more)*

Self-Hosted Deployment

Complete **On-premises** or **Private Cloud** deployment-data never leaves the customer environment

RAG-Powered Accuracy

Every response is grounded in actual documents with **full citation to sources**

Agentic Capabilities

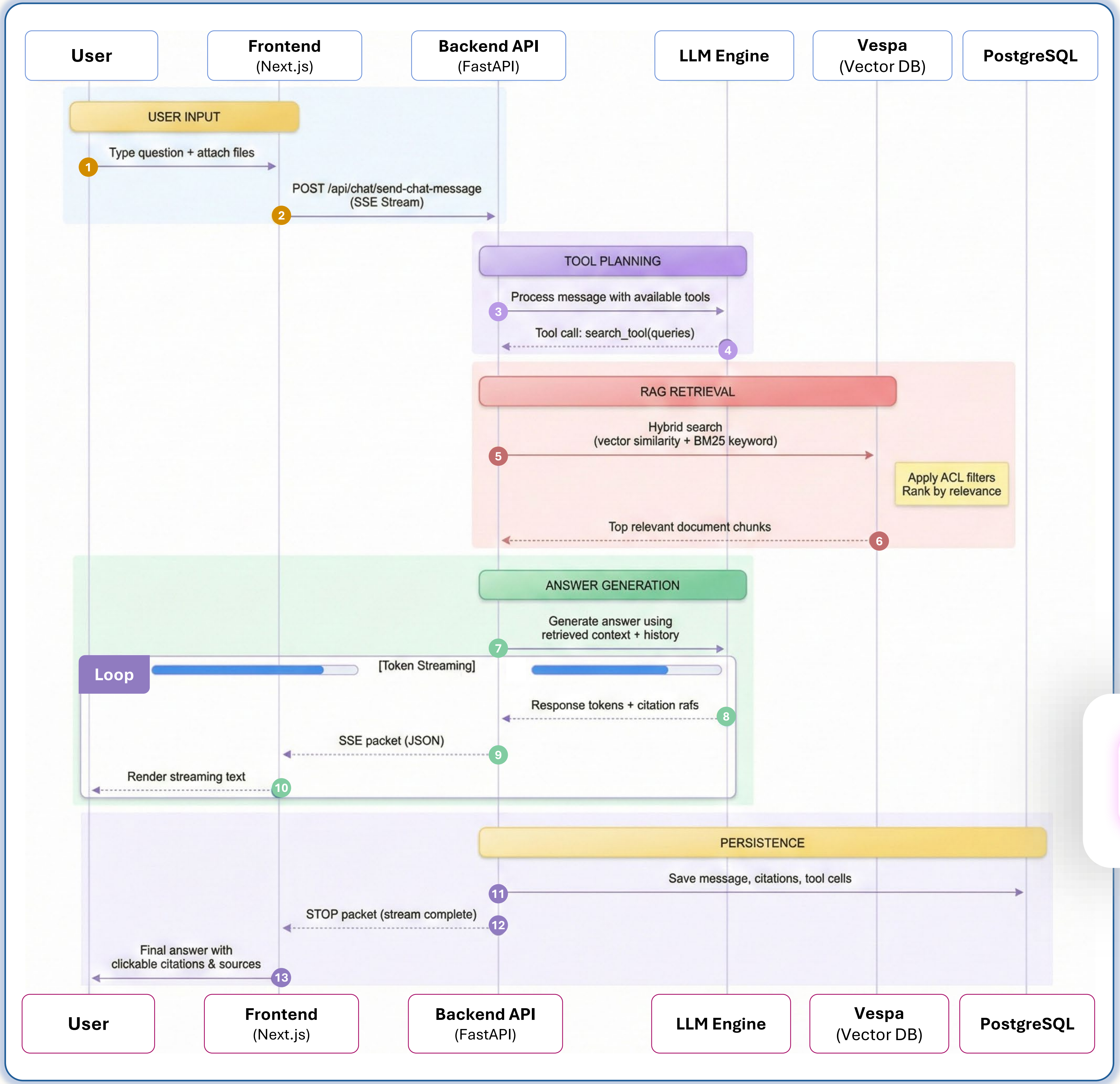
Deep research, code execution, web search, and custom tool integration

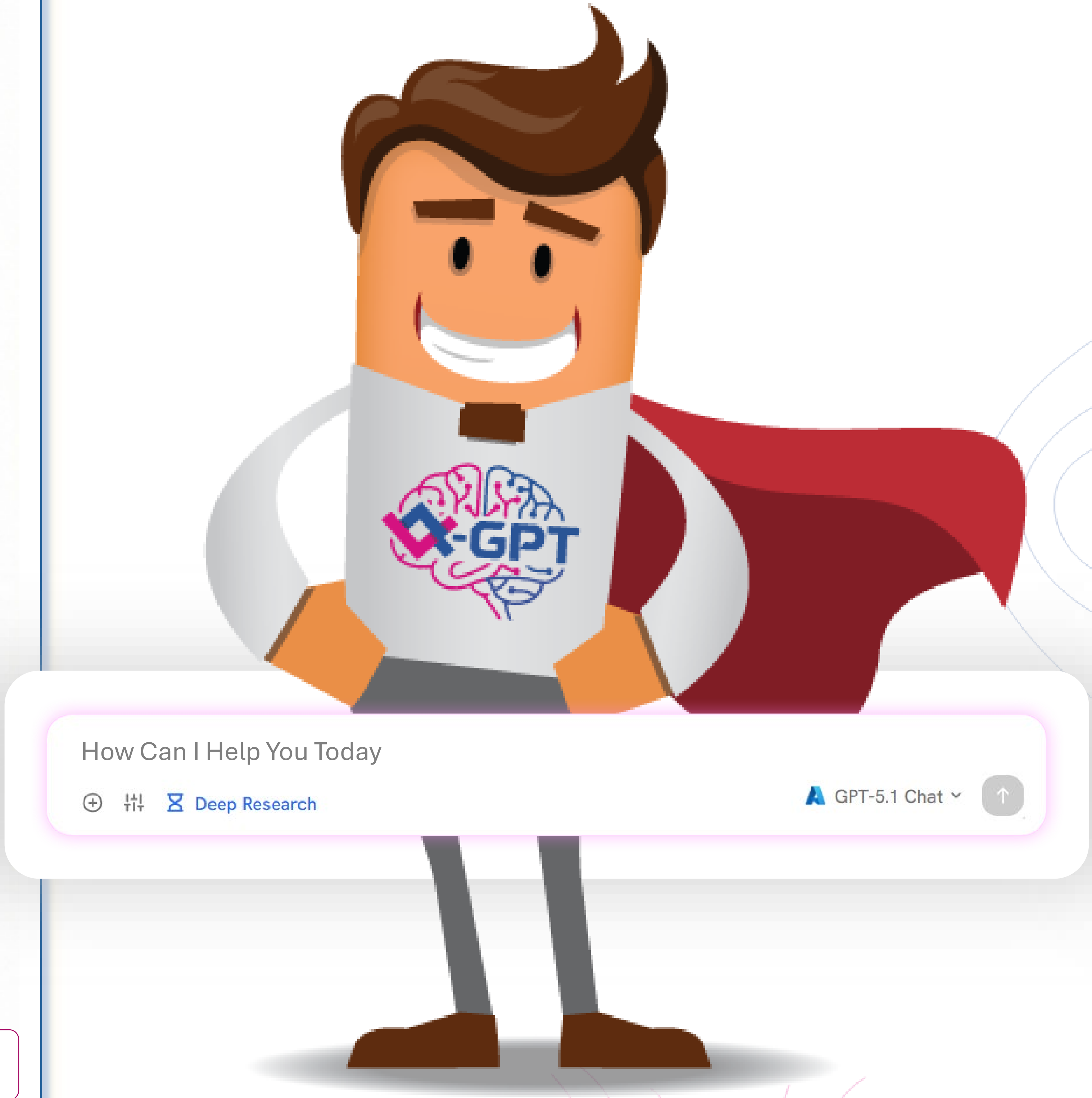
Permission Aware

Automatically syncs document permissions from source systems—users only see what they're authorized to access

https://qgpt.qbadvisory.com

Q-GPT Orchestrated AI Pipeline





https://qgpt.qbadvisory.com

Q-GPT - Built on Core AI Technologies

Large Language Models (LLMs)

- OpenAI (GPT-5, o3)
- Anthropic (Claude Opus 4.5)
- Google (Gemini 2.5)
- Local (Llama, Mistral, Qwen)

Retrieval Augmented Generation (RAG)

- Contextual Retrieval
- Intelligent Chunking
- Query Understanding
- StructRAG (Q2 2026)

Reranking

- Cohere Rerank
- AWS Bedrock
- LiteLLM-compatible rerankers
- Neural reranking improves top-result precision

Vector Database & Search

- Vespa (Primary)
- OpenSearch
- Hybrid Search (BM25 + Vector)

Embedding Models

- OpenAI text-embedding-3
- Cohere embed-v3
- Voyage AI
- Nomic embed-text (Local)

Agentic AI

- Deep Research
- Python Execution
- Web Search
- MCP Tools

Inference Optimization

- vLLM PagedAttention
- LiteLLM Universal Interface
- Real-time Streaming

Q-GPT Technology Innovation

Model Context Protocol (MCP) Integration

Q-GPT implements the emerging **MCP** standard, enabling interoperability with external AI clients and positioning the platform for the interconnected **AI ecosystem** of the future.

Hybrid Search Architecture

We combine **BM25** keyword search with dense vector similarity in a unified Vespa-powered index. **This hybrid approach captures** both semantic meaning and exact keyword matches, significantly improving retrieval accuracy over pure vector search.

vLLM High-Throughput Inference

For self-hosted deployments, **Q-GPT leverages vLLM's PagedAttention** technology, achieving **10-24x higher** throughput compared to **naive LLM serving**. This enables cost-efficient GPU utilization for enterprise-scale deployments.

Multi-Pass Contextual Retrieval

Our **RAG pipeline uses LLM-enhanced** chunk **contextualization**—adding document context to each chunk before embedding—dramatically improving retrieval relevance for ambiguous queries

Q-GPT – Leveraging Open-Source

Current AI Landscape

AI Learning Platforms

Sana 360Learning

- AI-enhanced LMS
- Training-focused
- SaaS-first deployment
- Limited cross-system intelligence

Enterprise AI Search

glean onyx

- Strong knowledge retrieval
- Cloud-native
- Closed model ecosystems
- Not infrastructure-level platforms

AI Automation Assistants

n8n Moveworks

- Workflow automation
- IT-centric use cases
- SaaS dependent
- Not learning-centric

Q-GPT Advantage

Sovereign by Design

- Fully self-hosted
- EU data compliance native
- Zero data externalization

Model Independence

- LLM-agnostic architecture
- No vendor lock-in
- Future-proof AI strategy

Enterprise-Wide Intelligence

- 40+ system connectors
- Hybrid + contextual retrieval
- Cross-application reasoning

AI Learning Everywhere

- Continuous learning across tools
- Not confined to LMS
- Embedded enterprise intelligence

https://qgpt.qbadvisory.com

qba

Q-GPT Product Roadmap

Our roadmap includes two breakthrough technologies that will further differentiate the platform

Q4 2025

Current Interfaces

Q1 2026

StructRAG with Knowledge Graph Integration

Q-GPT will go beyond traditional RAG by implementing structured knowledge representation based on the StructRAG research. The Knowledge Graph module will extract entities and relationships from documents, enabling:

- **Entity-Centric Queries**
- **Graph-Enhanced Retrieval**
- **Relationship Mapping**
- **Multi-Hop Reasoning**

Q2 2026

Recursive Language Models - RLM

A breakthrough inference strategy for handling unbounded context lengths. Unlike traditional approaches that struggle with "context rot" in long documents, RLM enables:

- **Context-Centric Decomposition**
- **Recursive Sub-Queries**
- **Unbounded Context Handling**
- **Smaller Models, Better Results**

Q-GPT - Role of AI

AI is not an add-on feature in Q-GPT—it is the core product. Every aspect of the platform is designed around AI capabilities

Accuracy Through RAG

Traditional chatbots hallucinate because they rely solely on LLM knowledge. Q-GPT grounds every response in actual retrieved documents, dramatically reducing hallucination rates, with a citation system that ensures users can verify any claim against source material.

LLM-Agnostic Architecture

Organizations have diverse needs-cost, latency, data residency. Q-GPT's LLM-agnostic architecture enables seamless provider switching: GPT-5 for complex reasoning, Claude for long documents, or self-hosted Llama for complete privacy.

Graph Intelligence

Simple RAG retrieves similar text chunks. Q-GPT's upcoming StructRAG (Q2 2026) understands entities and relationships, enabling structured reasoning across document boundaries for complex questions like which Q3 sales customers also appear in support tickets.

Permission Aware AI

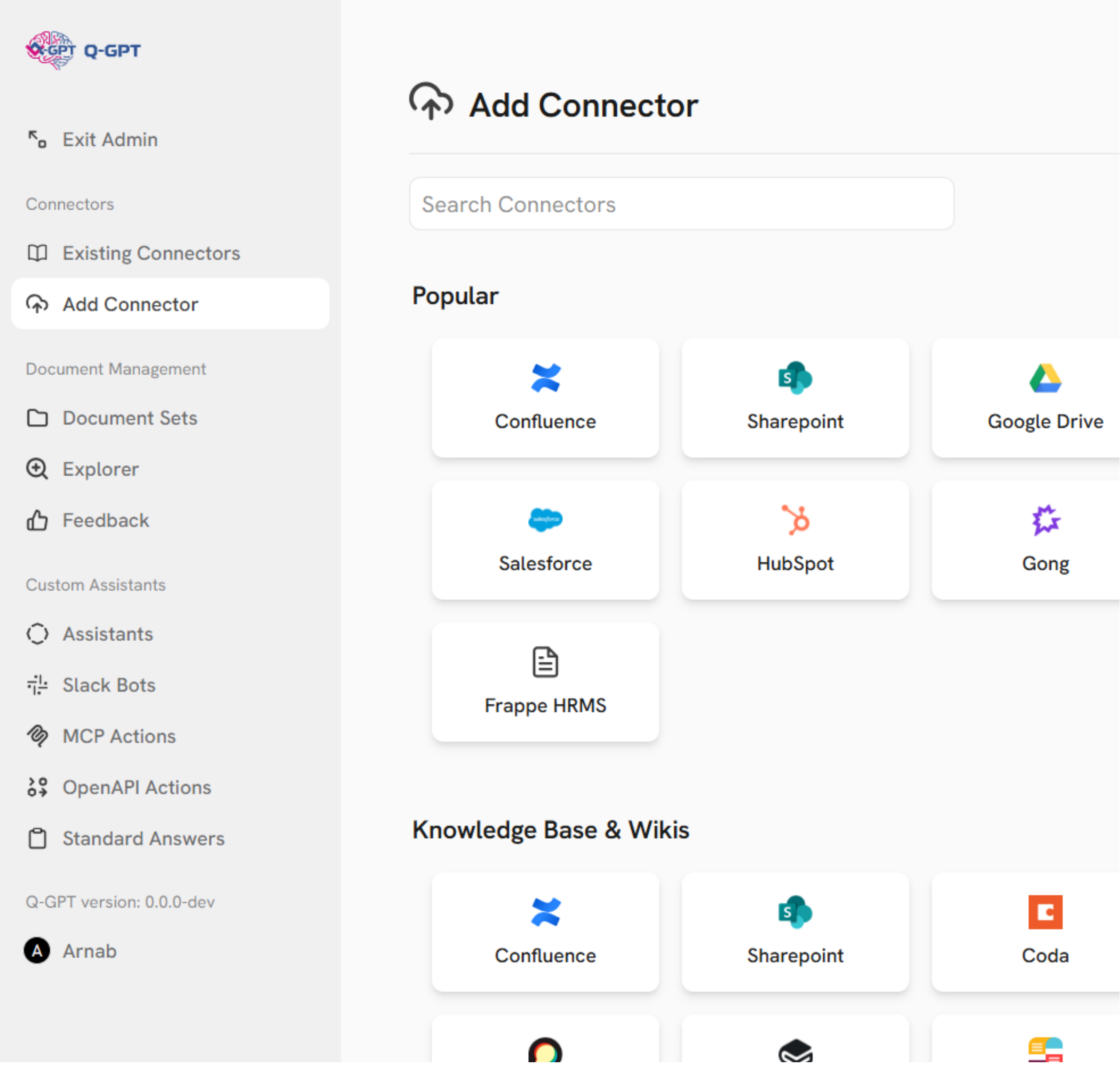
AI is only useful if it respects security boundaries. Q-GPT's automatic permission synchronization ensures the AI assistant can only access and surface information the user is authorized to see-maintaining security at scale.

Autonomy Through Agentic Features

Deep Research enables Q-GPT to autonomously plan and execute multi-step investigations - searching across sources, synthesizing findings, and delivering comprehensive reports in minutes instead of hours.

Q-GPT Configuration

Connector Management



Knowledge & Wikis	Confluence	Notion	SharePoint	Coda	GitBook
Cloud Storage	Google Drive	Dropbox	S3	Azure	Oracle
Collaboration	Slack	Teams	Gmail	Discord	Email
CRM & Sales	Salesforce	HubSpot	Gong Yoo	Fireflies	Highspot
Operations	Jira	Zendesk	Asana	Productboard	Freshdesk
Code	GitHub	GitLab	Bitbucket		

https://qgpt.qbadvisory.com

qba

Q-GPT Configuration

MCP Management

MCP Actions

Connect MCP (Model Context Protocol) servers to add custom actions and tools for your assistants.

Search servers...

Outlook
Arcade MCP Server

Teams
Arcade MCP Server

Google Search
Arcade MCP Server

Categories

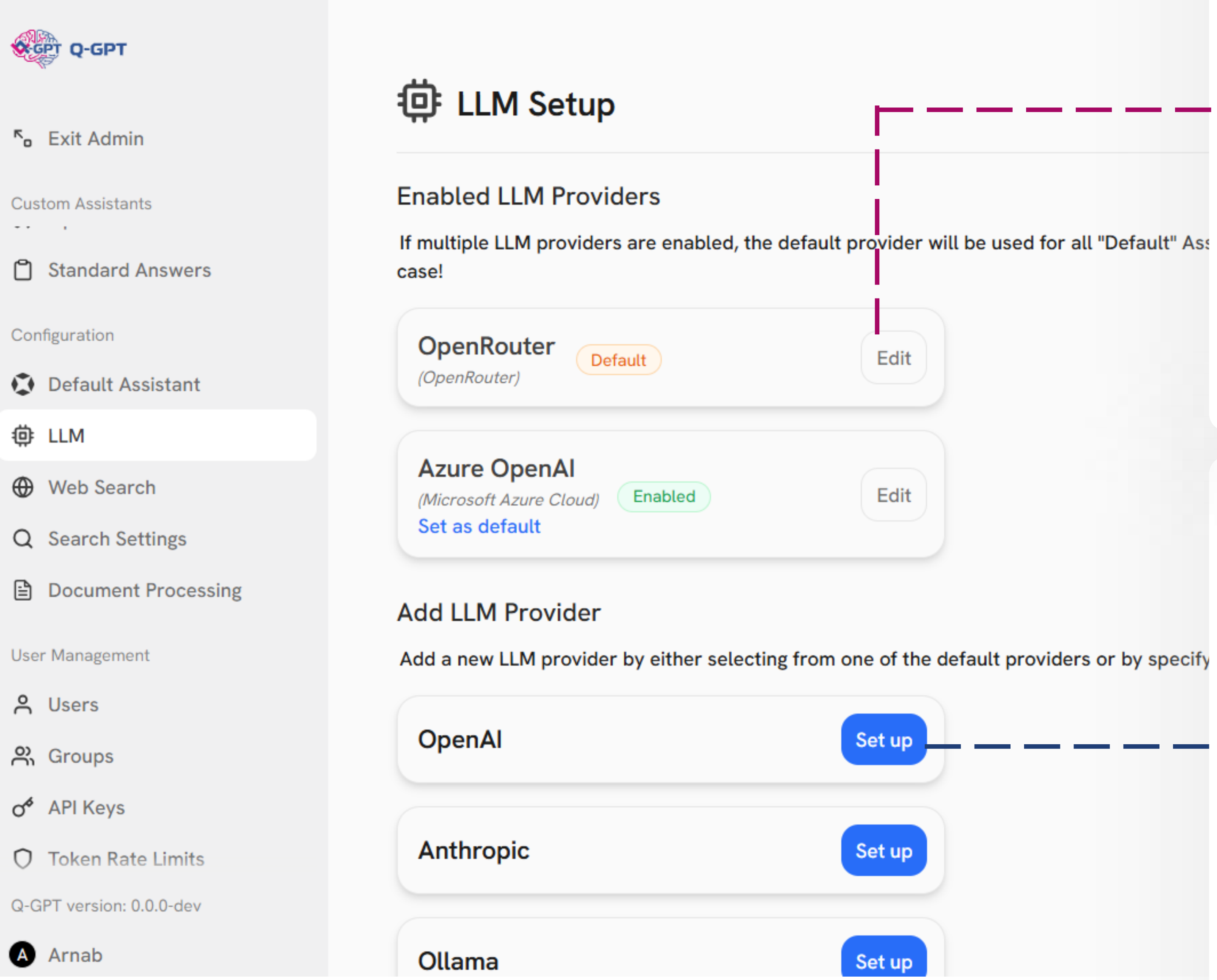
- All Tools
- Productivity & Docs
- Communication
- Developer Tools
- Search
- Sales & CRM
- Entertainment
- Finance & Payments
- Customer Support

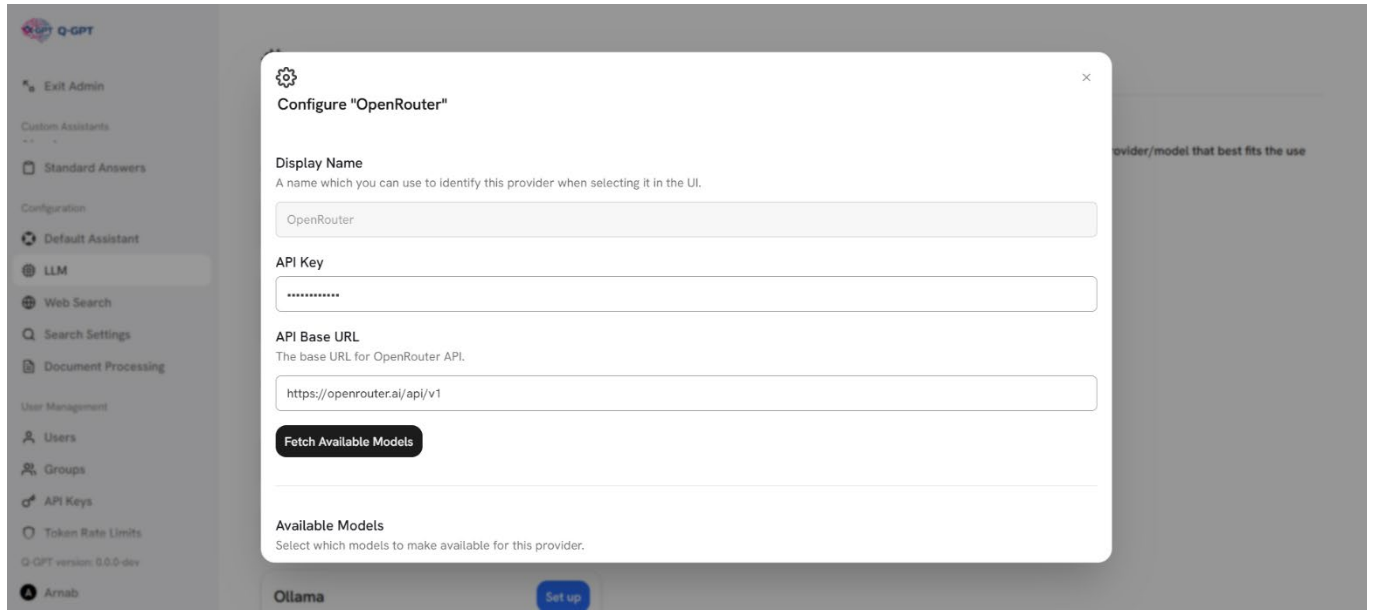
MCP Action Hub

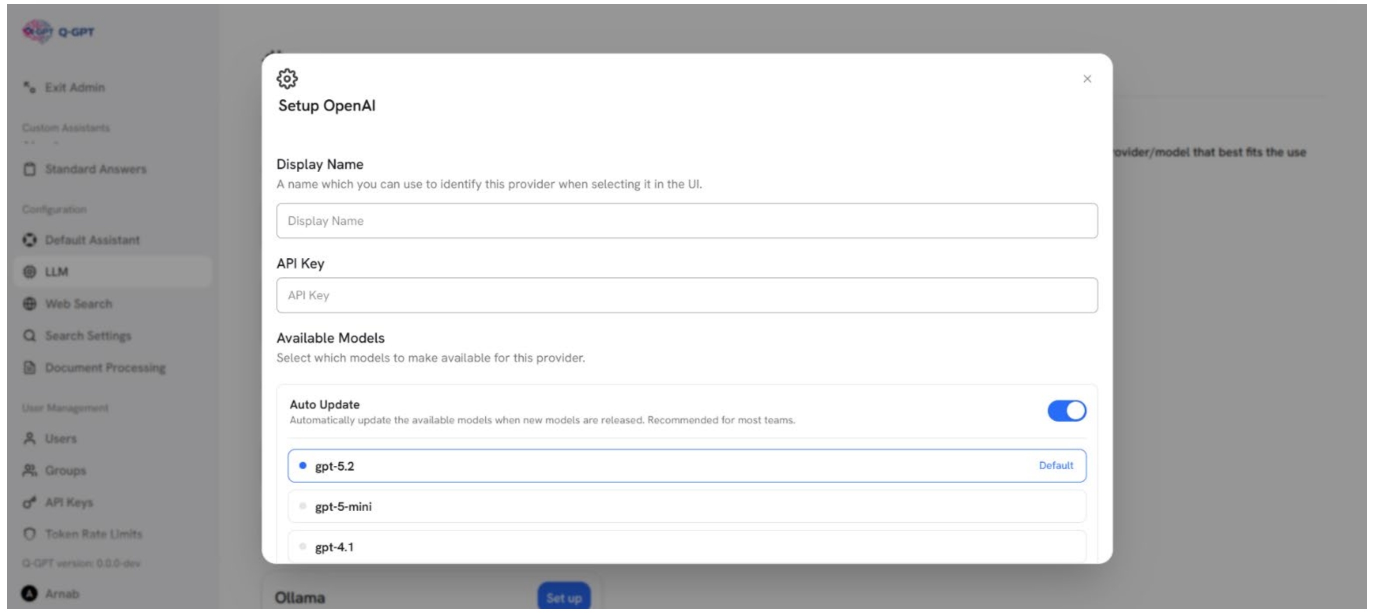
Confluence	SharePoint	Notion	Zendesk
Slack	Dropbox	Google Drive	Asana
Gmail	Teams	HubSpot	Salesforce
Asana	Outlook	Google Docs	Google Sheets

Q-GPT Configuration

LLM Setup







https://qgpt.qbadvisory.com

Q-GPT Metrics

Time Saved

up to 8 hours

Per Employee / Week

Query Response

< 2 sec

Average Time

Deploy Flexibility

On-prem/ Hybrid/ Private Cloud

Cost Reduction

60 – 80%

vs. Per-seat Solutions

Doc Coverage

40 +

Connector, Millions Indexed.

Open-Source Community Impact

Active contributor community enabling rapid feature development | Transparency in AI implementation—organizations can audit the code | Ecosystem of extensions, connectors, and integrations

Ethical Design & Safeguards within Q-GPT

Personal Data Handling

Data Sovereignty by Design
- Fully self-hosted — all data remains within customer infrastructure
- No external AI providers or third-party data sharing
- GDPR-compliant and aligned with EU data localization (ideal for regulated industries)

Enterprise-Grade Encryption
- AES-256 encryption at rest
- TLS 1.3 encryption in transit
- Encrypted backups and secure credential storage

Data Retention Controls
- Configurable TTL policies for chat history
- Automatic expiration and cleanup
- User-controlled conversation history

Transparency to Users

Citation System
- Every AI response includes source document citations
- Transparent visibility into referenced documents
- Enables verification and builds trust

Audit Logging
- Complete query and response history with timestamps
- User attribution for all interactions
- Admin action audit trails for compliance

Explainable Retrieval
- Visibility into retrieved documents
- Insight into document ranking logic
- Clear understanding of why specific answers were generated

Bias & Fairness Management

User Feedback Integration
- Users can upvote/downvote responses and documents
- Feedback adjusts document ranking scores over time
- Continuous improvement based on user signals

Source Diversity
- Configurable source filtering prevents over-reliance on single systems
- Time filtering reduces bias toward stale information
- Balanced retrieval across document types

Permission-Based Filtering
- Results are filtered based on user permissions
- Prevents information asymmetry from AI access
- Maintains existing organizational security boundaries

Risk & Incident Handling

Role-Based Access Control (RBAC)
- Granular roles
- SSO integration
- Multi-factor authentication support

Permission Synchronization
- Automatic sync from multiple source systems
- Post-query filtering to prevent unauthorized data exposure
- Conservative defaults -restrict access on sync failure

Incident Response
- Comprehensive logging for investigation
- Rate limiting to prevent abuse
- Query validation for security protection

Responsible Deployment
- Clear security best-practice documentation
- Regular open-source security updates
- Enterprise-grade support for critical deployments

Q-GPT is Compliant to EU AI Act (Regulation EU 2024/1689)

01 Risk Classification

- Limited / Minimal **Risk under EU AI Act**
- No prohibited practices (no manipulation, social scoring, biometric ID, emotion recognition, etc.)
- Enterprise knowledge retrieval tool — not critical infrastructure decision AI

02 Data Quality & Bias Mitigation (Article 10)

- RAG architecture using curated enterprise data (not open web)
- Bias monitoring via feedback loops
- Configurable source filtering & continuous refinement

03 Human Oversight (Article 14)

- Role-Based Access Control (RBAC) & admin controls
- Permission synchronization with organizational access rights
- No autonomous decision-making; feedback-driven improvement

04 Technical Robustness & Security (Article 15)

- AES-256 at rest, TLS 1.3 in transit
- Citation-based accuracy & verifiability
- Self-hosted, resilient, monitored infrastructure

05 GDPR & Data Protection

- Data minimization & purpose limitation
- Configurable TTL (storage limitation)
- No cross-border data transfer (self-hosted)
- Full organizational control for data subject rights

06 GPAI Compliance

- Multi-provider architecture with model transparency
- Provider selection based on compliance posture
- Ollama / vLLM support (open models)
- No training on customer data (retrieval, not fine-tuning)

07 Regulatory Sandbox Readiness

- Open-source, modular architecture
- Suitable for EU AI regulatory sandbox participation
- Enables collaboration on AI governance standards

08 Transparency (Article 52)

- Clear user awareness of AI interaction
- AI-generated responses with source citations
- Full audit trails (query history, user attribution, response logging)

Q-GPT - Built for European AI Excellence

Compliant by design • Transparent by default • Trustworthy by architecture

Q-GPT: Redefining Enterprise AI

Luxembourg-based enterprise AI aligned with **EU regulations** and values

Self-hosted, GDPR-compliant, and built for data sovereignty

Hybrid search, contextual retrieval, **vLLM** inference, and **agentic** capabilities

Production-grade platform with advanced roadmap (***StructRAG Q2, RLM Q3***)

Eliminates vendor **lock-in, privacy risks**, and **uncontrolled** AI exposure

Cost-efficient, enterprise-ready solution for **regulated industries**

MIT-licensed **open-source** model enabling **auditability** and **customization**

40+ native connectors with **automatic** permission **synchronization**

Thank You

Your Right-Sourcing IT Partner.

References

StructRAG:

Li, Z., et al. (2024). *StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-time Hybrid Information Structurization*. arXiv:2410.08815. https://arxiv.org/abs/2410.08815

RLM:

Zhang, A., et al. (2025). Recursive Language Models: Context-Centric Decomposition for Unbounded Context Handling. arXiv:2512.24601v1. https://arxiv.org/abs/2512.24601v1

EU AI Act:

European Commission. (2024). *Regulation (EU) 2024/1689 - Artificial Intelligence Act*. Official Journal of the European Union. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai