PRODYNA

Agentic AI Platform on Kubernetes

Motivation

For most organizations the challenge with generative and agentic AI is no longer prototyping, but operating it reliably at scale. Initial use cases are delivered quickly, yet each additional application adds complexity. Credentials are duplicated across teams, model consumption cannot be attributed to its source, every agent integrates its own tools, and deployment practices diverge from one project to the next. The result is a growing landscape of ungoverned integrations that security, finance, and platform teams are unable to oversee.

This offering establishes a governed platform in place of that fragmentation. A central AI gateway provides the single point of access to every model, externally managed or self hosted, and to governed MCP tool servers through the same gateway, ensuring that access control, budgets, and guardrails are enforced centrally rather than per project. New use cases reach production within days, each one governed and observable by default.

What We Bring

PRODYNA brings more than two decades of experience in enterprise software development, cloud native architecture, and Kubernetes operations for various customers across different industries. With this offering we deliver a governed, scalable Agentic AI Platform built on standard enterprise open source technologies:

- A central Open Source **AI Gateway** (agentgateway) for unified, OpenAI compatible access to externally managed and self hosted models — with **single sign-on** (OIDC/JWT), per group **access control** (RBAC), upstream credential injection, **token budgets**, guardrails, and per user/per agent cost attribution
- A governed **MCP gateway**: the same gateway fronts Model Context Protocol tool servers with shared authentication, per server authorization, and a self service registry
- A ready to use **chat UI** (LibreChat) with single sign-on (OIDC)
- Built-in **observability** with the Grafana stack: Prometheus metrics, Grafana Tempo (optional Langfuse) for **distributed LLM & agent tracing**, and ready made dashboards for usage, cost and budgets
- Optional **self-hosted LLMs** on Kubernetes (vLLM), optionally managed by KServe

What You Need

To make the best use of this offer and permit a fast and efficient start, you will need:

- A target environment (public cloud, private cloud, or on-premises) with infrastructure foundations in place
- Availability of your experts (e.g. platform, network, IAM, security)
- A small group of motivated IT colleagues to onboard as early adopters

Benefits

- **Portable by design:** Built entirely on Kubernetes, the platform runs in any environment, public cloud, private cloud, sovereign cloud, or on-premises, keeping your data and AI workloads under full control
- **Predictable costs:** Optionally self-host open-weight LLMs directly on Kubernetes. With only GPU capacity required, you avoid dynamic per-token pricing and gain full cost transparency
- **Self-service for developers:** A strict, opinionated "Golden Path" template lets your IT colleagues ship AI applications in days, without deep Kubernetes expertise
- **Governed from day one:** Central LLM gateway, access control (SSO + RBAC), cost tracking, guardrails, and built-in observability are inherited automatically by every agent

Quick facts

- **Duration:** ~ 12 weeks (depending on use case complexity and cluster setup)
- **Deliverables:** Agentic AI Platform with central gateway and observability, first AI agents, enablement for internal team

PRODYNA

What You Get

PRODYNA will guide you and your employees through the solution creation process in the following phases:

PHASE	ACTIONS
Kubernetes Setup (2-3 weeks)	• Provision a production-ready Kubernetes cluster (control plane, node pools, optional GPU nodes) • Configure ingress/gateway, secret management (vault), storage, and network policies • Establish baseline platform security and isolation model
AI Platform (MVP) (4-6 weeks)	• Deploy SSO and identity (OIDC): JWT authentication and per-group RBAC • Deploy the LLM Gateway (agentgateway): unified model access, upstream-credential injection, token budgets, guardrails, and per-user/per-agent cost tracking • Deploy the MCP gateway: governed Model Context Protocol tool servers with per-server authorization and a self-service registry • Deploy the chat UI (LibreChat) with SSO, per-user model menus and MCP tools in chat • Deploy Observability: Grafana Tempo (LLM & agent tracing) with the Prometheus Grafana stack (metrics, dashboards & logs) • Optional: self-hosted open-weight LLMs on GPU nodes with eg vLLM/SGLang (optionally via KServe)
GitOps Agent Deployment (3-4 weeks)	• Implement the "Golden Path" agent template (OpenAI-compatible service, agent.yaml) — every agent is automatically fronted by the gateway and inherits SSO, RBAC, budgets, guardrails and tracing • Set up CI/CD pipelines, container registry, and ArgoCD GitOps sync • Build the first agents and MCP tool servers side-by-side with early adopters • Knowledge transfer and onboarding documentation

Note: This is our standard methodology. We are happy to modify and adapt to the specific needs of your organization.

Get Started

To learn about pricing and how to get started, please contact info@prodyna.com.

About PRODYNA

PRODYNA is a technology consulting and software engineering company that helps organizations create business value through AI, cloud, data, and custom digital platforms, delivering solutions from strategy through implementation and operation.