MCP Gateway Checklist

truefoundry

This checklist is designed to help you evaluate the MCP Gateway or you may connect at anuraag@truefoundry.com for a 30-minute personalized walkthrough.

Sub-criteria	What should you evaluate?	Priority
MCP Server deployment /Packaging and hosting		
Containerization deployment of MCP Servers	Can you package MCP servers into containers and deploy them reliably on your infrastructure (Kubernetes or VMs)?	**Must have**
Blue/green or canary deploys	Can you roll out a new MCP server version with zero downtime by using blue/green or canary strategies? Confirm that readiness and liveness probes are configurable.	**Must have**
Autoscaling	Can the platform autoscale based on usage?	**Must have**
Time to first deploy	How long does it take to host a new MCP server?	**Depends on use case**
MCP Marketplace		
MCP Catalog & discovery	Does the Gateway add minimal latency overhead compared to direct API calls (e.g., <50ms P95 for time-to-first-token)?	**Nice to have**
Verification	Verify that each MCP Server provides clear name and descriptions of tools.	**Nice to have**
Example configurations for agent integration	Does the marketplace provide ready-to-use configuration snippets that allow quick integration of the tool with an agent or client?	**Nice to have**
Deploy from MCP Marketplace	Can teams publish a new MCP server or tool easily from the catalog?	**Nice to have**
Version history	Does the marketplace track MCP versions?	**Nice to have**
MCP Registry		
Self-service registration of hosted MCP servers	Can teams register MCP servers via an easy self-service form to with required details (name, description, URL) via UI, API, or GitOps?	**Must have**
Role Based Access Control (RBAC)	Should be able to define fine grained permissions by user, team, and application.	**Must have**
Supported transports	Can the registry accept HTTP and STDIO servers?	**Must have**
Supported Auth	Support multiple authentication modes (Header Auth, OAuth2, Dynamic Client Registration, platform-issued tokens) with secrets stored securely?	**Must have**
MCP Authorization and Authentication		
Single Sign-On (SSO)	Can users log in with your company's identity provider (Okta, Google Workspace, Azure AD, etc.) instead of creating new accounts?	**Must have**
OAuth2 support	Does the gateway support standard OAuth2 flows so that users and applications can authenticate securely using tokens issued by your identity provider?	**Must have**
Dynamic Client Registration (DCR)	Can the gateway automatically register itself (or MCP servers) with the identity provider to obtain client IDs and secrets, instead of requiring manual setup?	**Nice to have**

Header / API key authentication	Can the gateway work with MCP servers that still rely on simple API keys or header-based tokens?	**Must have**
Credential & token management	Are all secrets and tokens stored securely and rotated regularly?	**Must have**
Playground support		
Quick UI experimentation	Can users run experiments in an intuitive UI with configurable parameters (temperature, tokens, etc.) and immediately see model + tool responses?	**Must have**
LLM Switching	Does the Playground allow switching across multiple LLMs (OpenAI, Anthropic, Mistral, etc.) to experiment	**Must have**
Tool access with RBAC	Can users selectively enable MCP servers and tools in the Playground, with access governed by the same RBAC policies defined in the registry?	**Must have**
Prompt experimentation	Does the Playground allow users to write, test, and refine prompts interactively before deploying them into production pipelines or agents?	**Must have**
Guardrail validation	Can input/output guardrails (e.g., moderation checks, content filters) be applied directly in the Playground to validate their behavior quickly?	**Must have**
Test out quickly in UI	Can users run experiments in an intuitive UI with configurable parameters (temperature, tokens, etc.) and immediately see model + tool responses?	**Nice to have**
Code snippet generation	Does the Playground provide ready-to-use code snippets (Python, Node.js, REST, LangChain, LangGraph, etc.) that automatically include LLM, MCP server/tool integration and guardrail configuration?	**Must have**
Collaboration & sharing	Can users share Playground experiments (prompts + tool configs + parameters) with teammates?	**Must have**
Agent Registry		
Publish shareable agents from Playground	Can users publish tested agent apps directly from the Playground into agent registry with prompts, guardrails, and tools already bundled?	**Must have**
Discoverability	Does the registry provide a searchable catalog of published agent apps with metadata (name, description, owner, collaborators, version, usage)?	**Must have**
Access control (RBAC)	Can fine-grained permissions be applied to agent apps (e.g., who can view, who can manage)?	**Must have**
Versioning & updates	Does the registry support versioning of agent apps and allow rollback or safe upgrades when prompts, guardrails, or tools are updated?	**Nice to have**
Audit and Observability		
End-to-end request tracing	Can each request be traced across model and tool calls ?	**Must have**
Audit logs	Are all prompts, tool calls, responses, and admin actions logged with timestamps, actors?	**Must have**
Real-time Usage, cost & latency metrics	Does gateway expose per-user/team/application usage, cost, latency, and error rates?	**Must have**

OpenTelemetry compatibility	Can logs/metrics/traces from gateway be exported to OTEL/ Prometheus/Datadog/Splunk etc	**Must have**
Alerts & thresholds	Are SLOs, error spikes, or budget breaches monitored with configurable alerts?	**Nice to have**
Retention and privacy	Are retention policies for audit logs configurable with custom retention period and access?	**Nice to have**
Guardrails and Security		
Input/output guardrails	Are guardrails (moderation, regex, policy rules, PII masking) consistently enforced across both gateways?	**Must have**
AI-specific threat detection	Does the gateway detect and block prompt injection, response poisoning, data exfiltration, and jailbreaks (aligned to OWASP LLM Top 10)	**Must have**
MCP server security	Are MCP servers sandboxed with authentication, RBAC and network isolation?	**Must have**
Runtime security	Does the gateway enforce runtime hardening (signed containers, vulnerability scanning, mTLS, zero-trust networking)?	**Must have**
Rate limiting		
Per-user/team/app limits	Prevent abuse, DoS, or runaway agents consuming excessive resources.	**Must have**
Per-tool/server quotas	Are quotas configurable per MCP server or tool?	**Must have**
Fallbacks & retries	Does the system support configurable retries and failover?	**Must have**
Budget alerts	Can spend budgets be defined on a user/team or application level?	**Nice to have**
Ecosystem & Integration		
Multi-LLM providers	Can the gateway integrate with multiple model providers seamlessly (OpenAI, Anthropic, Mistral, etc.) using a unified interface?	**Must have**
Identity & SSO	Does it support enterprise IdPs (Okta, Azure AD, Google Workspace) with RBAC?	**Must have**
Observability & Evaluation tools	Can it integrate with external monitoring/evaluation systems (Prometheus, Datadog, Arize, Evals)?	**Must have**
CI/CD & Git	Are MCP configs, guardrails, and agent definitions version-controlled and deployable via GitOps or CI/CD?	**Must have**
Agent frameworks	Can MCP tools integrate with agent frameworks (LangChain, LangGraph, OpenAI Assistants)?	**Must have**
Runtime Security	Are MCP runtime environments isolated with sandboxing, egress control, and compliance guardrails?	**Must have**
Performance & Scalability		
Low latency	Does the Gateway add minimal latency overhead compared to direct API calls (e.g., <50ms p95 for time-to-first-token)?	**Must have**
High throughput	Can it handle a high volume of concurrent requests (e.g., 100+ RPS) without bottlenecks or degradation?	**Must have**
Horizontal scaling	Does it support autoscaling across multiple pods/nodes to meet load spikes?	**Must have**

Multi region hosting	Can the Gateway be deployed in multiple regions to support geographic redundancy and local latency optimization?	**Depends on use case**
Gateway Deployment & Configuration		
Deployment Options	Can the platform be deployed in self-hosted environments (VPC/ on-prem), managed cloud, or hybrid setups?	**Must have**
Terraform/Helm Support	Does it support infrastructure-as-code tools (e.g., Terraform, Helm) for repeatable deployments?	**Must have**
CI/CD Integration	Can the platform integrate with CI/CD pipelines to automate deployment and configuration updates?	**Nice to have**
Kubernetes Native Deployment	Is the platform deployable as a native Kubernetes workload (Helm charts, CRDs, operators, etc.)?	**Depends on use case**
Edge Deployment	Is the platform capable of being deployed at edge locations for latency-sensitive or offline scenarios?	**Depends on use case**
GitOps Workflow Support	Does it support GitOps-based configuration and deployment management?	**Nice to have**
Enterprise Considerations		
SLA & Support	Are SLAs available for uptime, support response times, and issue resolution?	**Must have**
Tenant/Workspace Isolation	Can you enforce strong isolation between teams, projects, or business units?	**Must have**
GDPR/SOC2/HIPAA Compliance	Does the platform offer controls or attestations for enterprise-grade compliance needs?	**Depends on use case**

About TrueFoundry

TrueFoundry is an enterprise platform-as-a-service that enables AI, data, and platform teams to build, deploy, and manage large language model applications at scale—with speed, cost efficiency, and robust governance. It streamlines the end-to-end AI lifecycle through capabilities like auto-scaling, proactive maintenance, and centralized access controls. Its AI Gateway provides a unified control layer across models and environments, ensuring reliability, observability, and security. Leading enterprises such as NVIDIA, Merck, Synopsys, and others trust TrueFoundry to accelerate innovation and deliver AI at scale. TrueFoundry's vision is to create a self-sustaining AI ecosystem where AI manages AI, driving unparalleled speed, scale, and innovation for businesses. To learn more about TrueFoundry, visit truefoundry.com.

Trusted by

NVIDIA, NetApp, MERCK, SIEMENS Healthineers, SYNOPSYS, FOXCONN, MAVENIR, innovaccer, AUTOMATION ANYWHERE, ResMed, zscaler, HOLCIM

Recognized by

Gartner, IDC Innovator 2024, HFS HOT TECH July 2024 truefoundry, Highly rated 9.9 G2, SUMMER 2025 Momentum Leader, SUMMER 2025 Best Est. ROI

Compliant and secure

AICPA SOC 2 Formerly SAS 70 Reports, HIPAA