

AI Gateway: matriz de features

Gateways proprietários com licenciamento comercial

PROPRIETÁRIOS / COMERCIAIS · licenciamento pago						
Feature	Conceito básico	Status	Kong AI Gateway Enterprise	Axway Amplify AI Gateway	MuleSoft AI Gateway	KrakenD Enterprise
ESSENCIAIS — base operacional sem a qual o gateway não cumpre seu papel						
Multi-LLM routing	Acessar OpenAI, Anthropic, Bedrock, Vertex e outros sob interface unificada.	MUST				
Auth e rate limiting	Autenticação (OAuth, API keys, mTLS) e controle de requisições por consumidor.	MUST				
Token-based rate limiting	Limites por tokens consumidos, não apenas por número de requisições.	MUST				
Cost tracking e budgets	Atribuição de custo por chave, time ou projeto, com orçamentos enforcáveis.	MUST				
Audit logs	Registro auditável de cada chamada para compliance e investigação.	MUST				
Observabilidade (OpenTelemetry)	Logs, métricas e traces compatíveis com a stack corporativa existente.	MUST				
IMPORTANTES — definem se o gateway atende produção corporativa real						
Semantic caching	Cache de respostas baseado em similaridade semântica do prompt, não match exato.	IMP				
PII protection e data masking	Sanitização de dados sensíveis em prompts e respostas antes de chegar ao LLM.	IMP				
Prompt injection detection	Deteção de ataques de injeção no prompt antes de ir para o modelo.	IMP				
MCP server hosting	Capacidade de expor serviços internos como MCP tools consumíveis por agentes.	IMP				
MCP gateway e federação	Federar múltiplos MCP servers sob endpoint único com governança centralizada.	IMP				
Agent-to-agent (A2A)	Suporte ao protocolo de comunicação direta entre agentes autônomos.	IMP				
Multi-tenancy hierárquica	Estrutura Organização → Time → Usuário → Chave com herança de políticas.	IMP				
DIFERENCIAIS — onde os produtos se destacam ou se afastam do pacote						
Semantic routing	Rotear o prompt para o modelo mais adequado com base no conteúdo da requisição.	NICE				
RAG injection	Injeção automática de contexto vindo de vector store antes do prompt chegar ao LLM.	NICE				
Federated multi-cluster	Operar sobre gateways de terceiros (AWS, Azure, Istio) sem substituí-los.	NICE				
Marketplace de MCP servers	Catálogo interno onde áreas publicam e consomem MCP tools com governança.	NICE				
Prompt templates e decoration	Padronização de prompts entre times com templates reutilizáveis.	NICE				

Suporte nativo Parcial, plugin ou em desenvolvimento Não suporta

MUST Essencial: gateway que não tem isso não cumpre seu papel **IMP** Importante: define se o gateway entra em produção corporativa **NICE** Diferencial: onde os produtos se afastam do pacote

Cada produto tem sua base instalada. Diferenças reais estão em MCP, A2A e federação.

Para conta com Axway B2B/MFT, integração vertical vence. Para conta cloud-native, Kong vence.

AI Gateway: matriz de features

Gateways open-source com licenças Apache 2.0, MIT e Linux Foundation

OPEN-SOURCE · licenças permissivas, sem custo de licenciamento						
Feature	Conceito básico	Status	APISIX Apache 2.0	LiteLLM MIT	ContextForge Apache 2.0	Solo.io agentgateway
ESSENCIAIS — base operacional sem a qual o gateway não cumpre seu papel						
Multi-LLM routing	Acessar OpenAI, Anthropic, Bedrock, Vertex e outros sob interface unificada.	MUST	●	●	●	●
Auth e rate limiting	Autenticação (OAuth, API keys, mTLS) e controle de requisições por consumidor.	MUST	●	●	●	●
Token-based rate limiting	Limites por tokens consumidos, não apenas por número de requisições.	MUST	●	●	●	●
Cost tracking e budgets	Atribuição de custo por chave, time ou projeto, com orçamentos enforcáveis.	MUST	●	●	◐	◐
Audit logs	Registro auditável de cada chamada para compliance e investigação.	MUST	●	●	●	●
Observabilidade (OpenTelemetry)	Logs, métricas e traces compatíveis com a stack corporativa existente.	MUST	●	●	●	●
IMPORTANTES — definem se o gateway atende produção corporativa real						
Semantic caching	Cache de respostas baseado em similaridade semântica do prompt, não match exato.	IMP	○	◐	◐	●
PII protection e data masking	Sanitização de dados sensíveis em prompts e respostas antes de chegar ao LLM.	IMP	●	◐	●	●
Prompt injection detection	Deteção de ataques de injeção no prompt antes de ir para o modelo.	IMP	◐	◐	◐	●
MCP server hosting	Capacidade de expor serviços internos como MCP tools consumíveis por agentes.	IMP	◐	◐	●	●
MCP gateway e federação	Federar múltiplos MCP servers sob endpoint único com governança centralizada.	IMP	◐	◐	●	●
Agent-to-agent (A2A)	Suporte ao protocolo de comunicação direta entre agentes autônomos.	IMP	○	●	●	●
Multi-tenancy hierárquica	Estrutura Organização → Time → Usuário → Chave com herança de políticas.	IMP	◐	●	◐	◐
DIFERENCIAIS — onde os produtos se destacam ou se afastam do pacote						
Semantic routing	Rotear o prompt para o modelo mais adequado com base no conteúdo da requisição.	NICE	◐	◐	◐	◐
RAG injection	Injeção automática de contexto vindo de vector store antes do prompt chegar ao LLM.	NICE	●	◐	◐	◐
Federated multi-cluster	Operar sobre gateways de terceiros (AWS, Azure, Istio) sem substituí-los.	NICE	◐	◐	●	●
Marketplace de MCP servers	Catálogo interno onde áreas publicam e consomem MCP tools com governança.	NICE	○	○	●	◐
Prompt templates e decoration	Padronização de prompts entre times com templates reutilizáveis.	NICE	●	◐	◐	◐
● Suporte nativo ◐ Parcial, plugin ou em desenvolvimento ○ Não suporta MUST Essencial: gateway que não tem isso não cumpre seu papel IMP Importante: define se o gateway entra em produção corporativa NICE Diferencial: onde os produtos se afastam do pacote						
APISIX cobre o essencial. ContextForge e Solo.io lideram em MCP e A2A. LiteLLM domina LLM proxy. Nenhum OSS puro cobre B2B/MFT legado. Essa camada exige integração adicional acima do gateway.						

AI chat corporativo: 3 categorias, 12 produtos

Essenciais universais + condicionais que dependem do caso + diferenciais

			OSS SELF-HOSTED				SAAS LLM ENTERPRISE				PLATAFORMAS CORPORATIVAS AI			
Feature	Conceito básico	Status	Open WebUI OSS	LibreChat OSS	AnythingLLM OSS	Lobe Chat OSS	ChatGPT OpenAI	Claude Anthropic	Copilot 365 Microsoft	Gemini Google	Amazon Q AWS	Glean independente	Copilot Studio Microsoft	Onyx open-core
ESSENCIAIS UNIVERSAIS -- sem estes, o produto não passa em auditoria corporativa														
SSO / SAML / OIDC	Autenticação corporativa com identidade única do funcionário.	MUST	●	●	◐	◐	●	●	●	●	●	●	●	●
Multi-user com RBAC	Papéis diferenciados por função, com permissões granulares.	MUST	●	◐	●	◐	●	●	●	●	●	●	●	●
Data residency	Dados corporativos não saem da rede ou região definida (LGPD, GDPR, BACEN).	MUST	●	●	●	●	●	●	●	●	●	●	●	●
Audit logs	Registro auditável de conversas e ações para compliance e resposta a incidente.	MUST	●	●	◐	◐	●	●	●	●	●	●	●	●
CONDICIONAIS — viram must-have conforme setor, arquitetura e caso de uso														
Compliance certs (SOC 2, HIPAA)	Must-have para banco, saúde, seguros, governo. Nice-to-have para outros setores.	COND	○	○	○	○	●	●	●	●	●	●	●	●
Deploy on-premise / self-hosted	Must-have se dados regulados por localização física (defesa, saúde específica).	COND	●	●	●	●	○	○	○	○	○	◐	○	●
OpenAI-compatible endpoint	Must-have se estratégia é AI Gateway próprio governando tráfego LLM.	COND	●	●	●	●	○	○	○	○	○	○	○	●
Multi-provider LLM	Must-have se estratégia é neutralidade de provedor (Claude + GPT + Ollama).	COND	●	●	●	●	○	○	○	○	◐	●	◐	●
RAG sobre documentos internos	Must-have se caso de uso é chat sobre base de conhecimento corporativa.	COND	●	●	●	●	●	●	●	●	●	●	●	●
Conectores corporativos	Must-have se caso de uso exige integração com SaaS (Confluence, Jira, SharePoint).	COND	◐	◐	◐	◐	◐	◐	●	●	●	●	●	●
DIFERENCIAIS — otimização, adoção ou operação em escala														
Workspaces multi-tenant	Áreas ou times isolados com suas próprias configurações e dados.	NICE	●	◐	●	◐	●	●	●	●	●	●	●	●
MCP tools nativo	Suporte ao protocolo MCP para conectar ferramentas corporativas.	NICE	●	●	●	●	●	●	◐	◐	◐	◐	●	◐
LDAP / Active Directory	Integração com diretório corporativo legado.	NICE	●	●	◐	○	●	●	●	●	●	●	●	●
SCIM 2.0 provisioning	Provisionamento automático de usuários e grupos em escala grande.	NICE	●	◐	○	○	●	●	●	●	●	●	●	●
Custom branding / white-label	Logo, cores e nome da empresa aplicados na interface.	NICE	◐	●	◐	●	◐	◐	◐	◐	◐	●	●	●
Air-gapped deployment	Deploy 100% isolado, sem qualquer conexão com internet (defesa, gov).	NICE	●	●	●	●	○	○	○	○	○	○	◐	●
Programação de tarefas	Executar prompts ou agentes de forma agendada (cron, recurring).	NICE	●	◐	◐	●	●	◐	●	●	●	●	●	◐
Controle remoto via API	Administrar plataforma e disparar interações via API externa.	NICE	●	●	●	●	●	●	●	●	●	●	●	●

● Suporte nativo ◐ Parcial, em desenvolvimento ou versão paga ○ Não suporta

MUST Universal: sem isso o produto não passa em auditoria **NICE** Condicional: vira must-have por setor, arquitetura ou caso de uso **NICE** Diferencial: otimização, adoção ou operação em escala

Os 4 essenciais universais são baseline: todos os 12 produtos cumprem em nível aceitável.

A decisão real está nos condicionais e diferenciais. Depende de setor regulado, estratégia de LLM e caso de uso do cliente. Não existe produto universalmente melhor.