

Enterprise Generative AI Platform on AWS

— Reference Implementation

An evolving reference architecture for secure, private and reusable AI workloads using Amazon Bedrock and EKS.

**Secure**

- OAuth 2.0
- Cognito authentication
- Private workloads
- No public access to EKS

**Scalable**

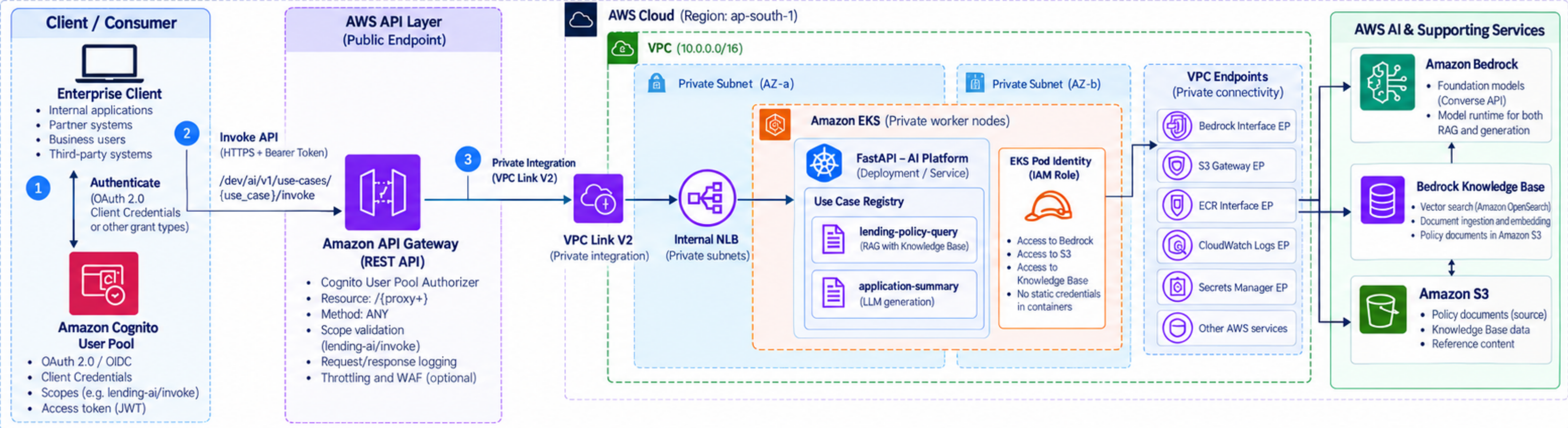
- EKS with auto scaling
- Modular use-case design
- Managed AWS services

**Enterprise-ready**

- Security, observability governance and multi-tenant patterns

**Reusable**

- Generic use-case platform
- Easy to add new AI workflows



1 Authenticate

Client authenticates with Amazon Cognito and receives an access token with the required scope.

2 Invoke API

Client calls the API Gateway REST API with the access token in the Authorization header.

3 Authorize

API Gateway validates the token and required scope using Cognito User Pool Authorizer. Requests without a valid token are rejected (401).

4 Private Integration

API Gateway uses VPC Link V2 to securely forward requests to an internal NLB in private subnets.

5 Route to EKS

NLB routes traffic to the FastAPI service running on Amazon EKS (private subnets, no public endpoint).

6 Execute Use Case

FastAPI routes the request to the appropriate use case (RAG or generation) and invokes Amazon Bedrock. Additional use cases can be added through the registry pattern.

7 Return Response

The platform returns a standardized response with the generated answer and source references (for RAG use cases).

Development & CI/CD

- GitHub Actions:** Build and test container image, Push to Amazon ECR (OIDC), Deploy to Amazon EKS, No long-lived AWS credentials.
- Amazon ECR:** Container image repository, Immutable image tags, Scan for vulnerabilities (optional).

Observability (Planned/Enhancement)

- Amazon CloudWatch:** Application logs, Metrics and alarms, Container insights.
- AWS X-Ray:** Tracing and latency analysis, End-to-end request tracing, Troubleshooting.
- Bedrock Model / Usage Metrics:** Token usage monitoring, Cost tracking, Error detection and alerts.

Security & Governance (Planned/Enhancement)

- AWS IAM:** Least privilege access, EKS Pod Identity, OIDC federation for CI/CD.
- AWS WAF:** API protection (optional), Rate limiting, Threat protection.
- Bedrock Guardrails:** Content safety, Prompt injection defense, Policy-based controls.

+ Future Extensions

- Additional AI use cases through registry
- Multi-tenant document access
- Per-use-case authorization and scopes
- Advanced evaluation and monitoring
- Full production hardening and resilience