TinyLlama Architecture - Intermediate Phase (2025-06)

CI/CD & Ops

- CloudWatch (metrics, alarms)
- SSM RunCommand
- AWS Budgets €15 warn / €20 stop
- Secrets Manager
- CodePipeline
- CodeBuild (unit tests ≥90%)
- ImageBuilder (Build AMI)

Compliance check
Alarm (via Lambda)
OpenAI key
Trigger build
Log builds/tests
Build AMI
Deploy AMI

Desktop

- Desktop GUI

Edge/API [EPIC: API]

- CloudFront (optional)
- Cognito JWT
- API Gateway HTTP API

HTTPS (w/ JWT)
Authorizer + API call

Router & Queue

- Lambda Router v2 (EPIC: Lambda)
- ElastiCache Redis (5m TTL, SM)

Response
Enqueue job

GPU Node & Storage

- EC2 g4dn.xlarge (vLLM, AMI)
- KMS-CMK (EBS/S3 encrypt)
- NAT Gateway
- VPC Private
- 100 GB gp3 cache
- S3 tinyllama-models

HTTPS egress
Private subnet
Data (boot/cache)
Model download
Upload inference result
Encrypt
Encrypt