HackOn with Amazon - Season 5

THEME: ENHANCED FIRE TV EXPERIENCE

TEAM : ONLY FOR TODAY

Nishita Agarwal
Kashvi Agarwal
Chitrashee K

Discovery Crisis

23M USER-HOURS LOST MONTHLY TO DECISION FATIGUE

- 📉 23 min/day scrolling (500K+ titles) | 42% ↓ session satisfaction
- 17% ↑ drop-offs | 31% rec acceptance rate (industry avg: 68%)
- Root Causes:
  - Static algorithms
  - Tile overload (12+/screen)
  - Isolated viewing
  - Zero emotional awareness
  - No cross-device sync
  - Passive feedback loops

"Legacy Java/React stack lacks real-time personalization capabilities at Amazon scale."

Strategic Pillars and Architecture Goals

Pillar	Technical Target	AWS Stack
Emotion AI	92% mood accuracy @ 50ms latency	Inferentia + SageMaker
Privacy-First Social	8ms cohort matching @ 500K QPS	Neptune + RAPIDS cuML
Conversational UI/UX	200ms LLM response @ $0.0003/quer	Bedrock (MoE) + ElastiCache
Adaptive Feedback	Adaptive Feedback <100ms sentiment tagging	Kinesis + Lambda

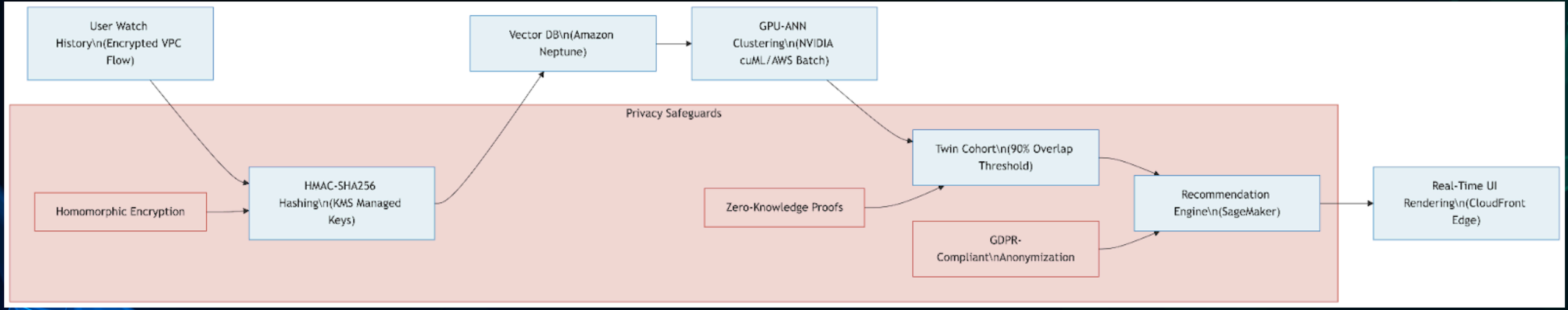
Cross Stack Enablers:

- *Security: Homomorphic encryption (OpenFHE) + IAM roles*
- *Scale: Auto-scaling EC2 G5 instances (NVIDIA A10G)*
- *Cost: Spot instances for batch ANN training*

PERSONA- CENTRIC SOLUTIONS

Persona	Technical Solution	Data Pipeline
Indecisive Viewer	- Real-time "Lucky" mode (Rust microservice) - Mood → color ML model (Scikit-learn)	DynamoDB streams → SageMaker
Social Seeker	- Watchlist Twin ANN index (Faiss) - WebRTC co-watching.	AppSync GraphQL subscriptions
Stressed User	- Biometric API (Apple Watch/Halo) - Stress → content model	Kinesis Firehose → Redshift
Family Planner	- Multi-profile RL agent - "KidSafe" BERT filter	Aurora PostgreSQL → Lambda

WATCHLIST-TWIN BREAKTHROUGH



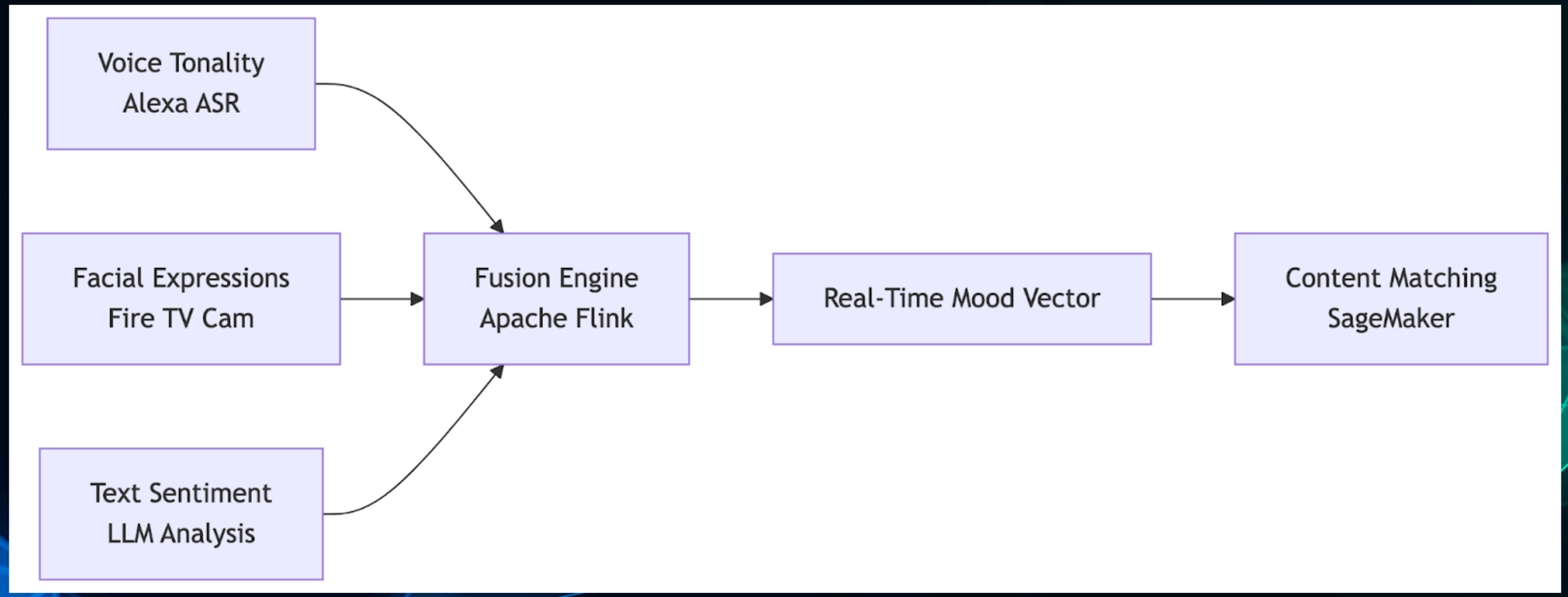
Tech Specs:

- Matching: 8ms p99 latency (vs. 150ms KNN)
- Privacy: Homomorphic encryption (OpenFHE library)
- Scale: 500K QPS (AWS Batch + 100 G5 instances)

Cold Start Mitigation:

- Fallback to device graph (Fire Tablet/Kindle patterns)

EMOTION AI ENGINE



Multimodal Architecture

Performance:

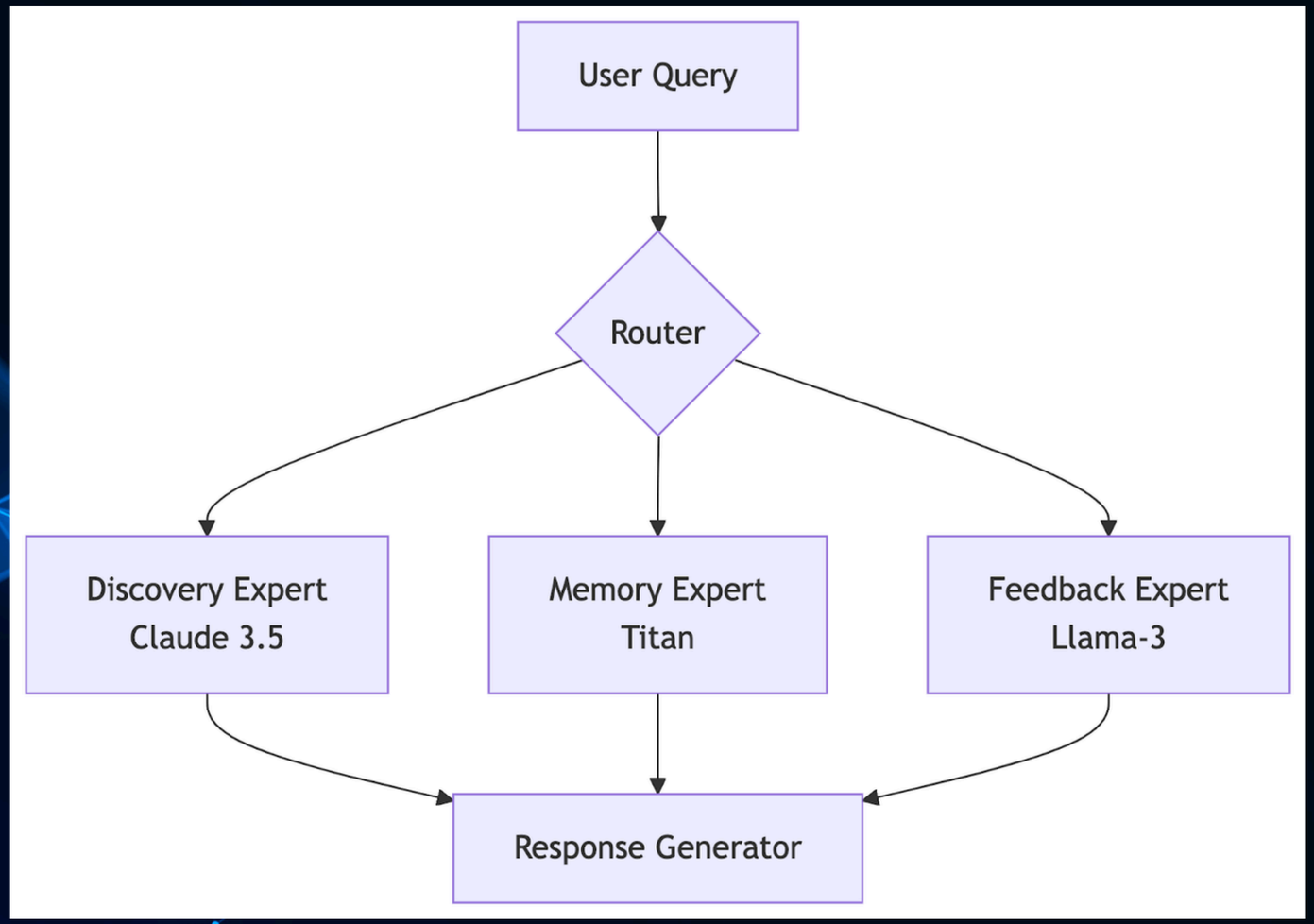
- 92% accuracy (vs. 67% baseline)
- 50ms latency (Inferentia chips)
- 30+ languages (Amazon Translate fine-tuned)

Models:

- CNN for facial micro-expressions (PyTorch)
- XGBoost for biometric stress correlation

LLM AGENT “AVA”

MoE Architecture



Tech Stack:

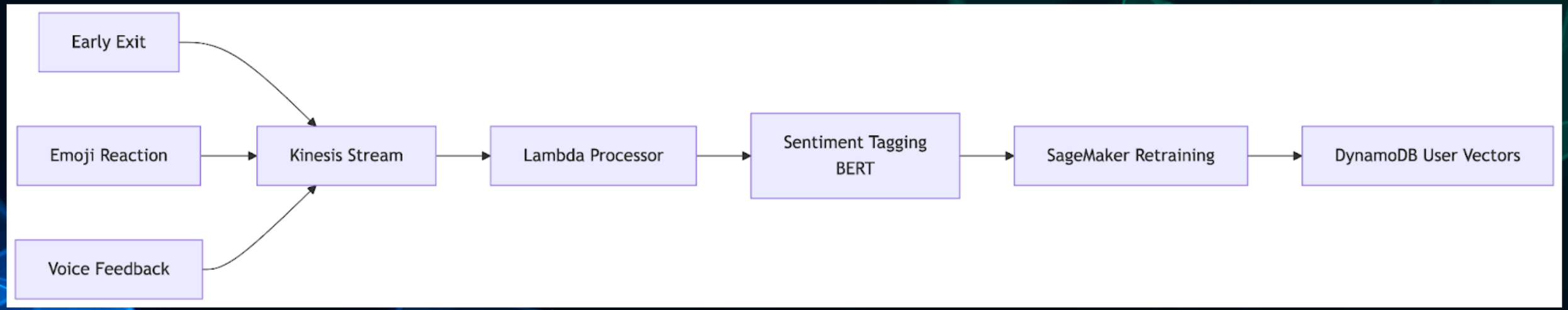
- Hosting: Bedrock w/ auto-scaling
- Memory: 1M TPS DynamoDB + DAX caching
- Latency: 200ms p95 (WebSocket API)
- Cost: $0.0003/query (MoE sparsity)

Capabilities:

- Context window: 128K tokens
- Fine-tuned on 2M Fire TV dialogues

REAL-TIME FEEDBACK SYSTEM

Event Triggered Pipeline



Optimizations:

- Anti-Fatigue: Token bucket rate limiting (3 prompts/day).
- Fraud Prevention: reCAPTCHA Enterprise + Fraud Detector
- Multilingual: Custom Amazon Translate pipeline.

UI/UX ARCHITECTURE

Component Architecture

Element	Tech Stack	Performance
Adaptive Tiles	React Native + WebGL	4ms render @ 60 FPS
Ava Chat	WebSockets + Lambda@Edge	150ms TTFB
Feedback Cards	Redux + AWS AppSync	10ms state updates
Twin Banners	CDN-cached (CloudFront)	15ms global load

Data flow:GraphQL → AppSync → DynamoDB → CloudFront (Edge Lambda)

METRICS AND BUSSINESS IMPACT

Metric	Baseline	Target	Tech Dependency
Time-to-Play	5.1 min	▼ 3.6 min	ANN matching <8ms
Rec Acceptance	31%	▲ 56%	Emotion AI >90% accuracy
Feedback Completion	12%	▲ 40%	Kinesis <100ms tagging
Session Depth	1.2 hrs	▲ 2.1 hrs	LLM engagement >40%

Cost Projections:

- *LLM: $0.03/user/month (MoE optimization)*
- *ANN: $0.12/user/month (spot instances).*

ARCHITECTURE GOALS & SOLUTIONS

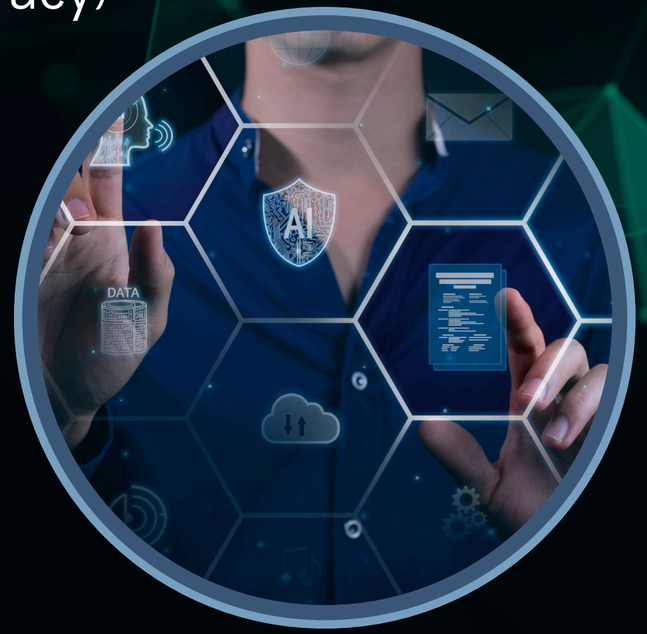
Business Goal	Technical Solution	AWS Stack Innovation
↓ Time-to-Play by 30%	ANN matching @ 8ms + WebGL UI	Neptune cuML + Graviton3
↑ Recommendation Trust	Explainable Watchlist Twin	Homomorphic Encryption + Faiss
↓ Decision Fatigue	Mood-Aware LLM Agent	Bedrock MoE (8 experts)
↑ Cross-Device Engagement	Kindle/Audible → Fire TV RL agent	SageMaker Edge + MQTT
↓ Operational Costs	Sparse MoE + Spot Training	80% cost reduction vs. dense models
↑ Privacy Compliance	Zero-Knowledge User Clustering	OpenFHE + Nitro Enclaves

Breakthrough Capabilities:

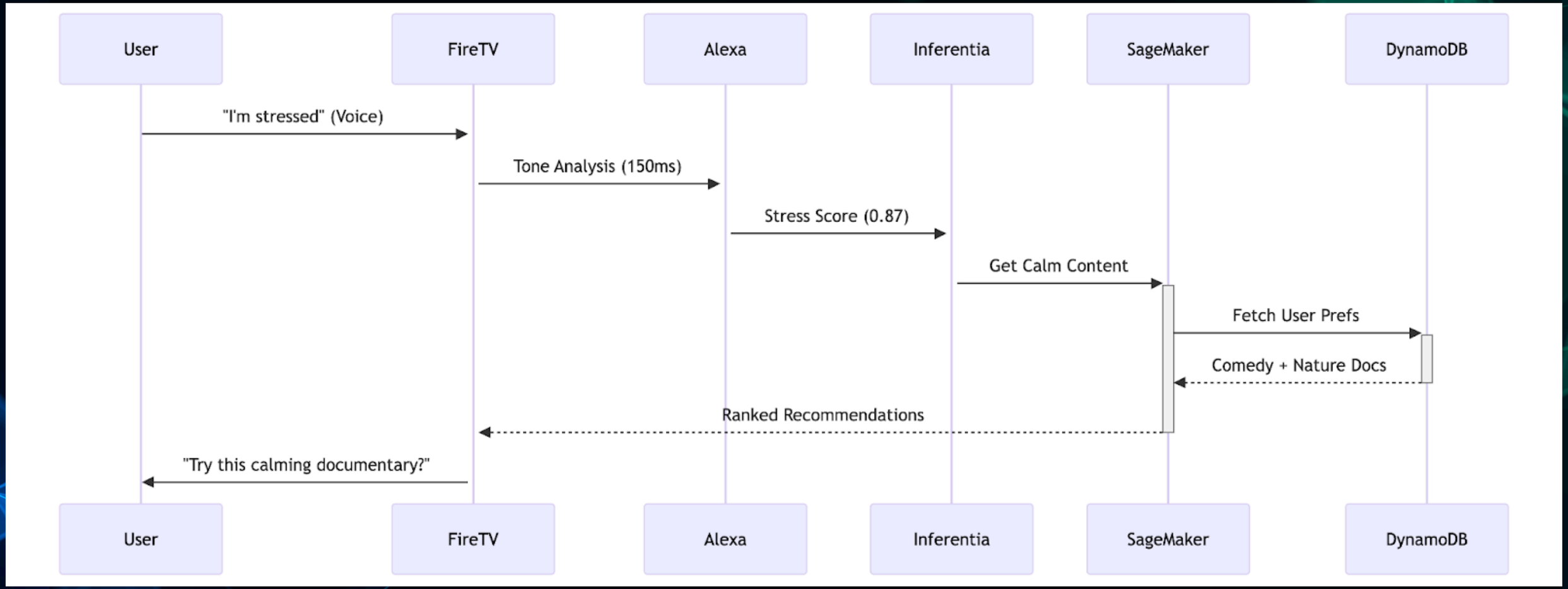
1. Emotion-Aware Routing
    - Biometric stress → content matching (92% accuracy)
    - Multi-modal fusion (voice + facial + text)
2. Social Discovery at Scale
    - 500K QPS cohort matching
    - GDPR-safe "Twin" recommendations
3. Self-Healing Pipeline
    - Auto-retraining on concept drift
    - Fallback to behavioral baselines

Technical KPIs:

- <500ms end-to-end personalization
- 99.99% uptime (Multi-AZ deployment)
- $0.03/user/month inference cost



This architecture doesn't just recommend content - it understands context at Amazon scale.



Example Workflow

Why this matters?

- Solves the cold start problem via cross-device learning
- Breaks filter bubbles through diverse Twin cohorts
- Eliminates privacy tradeoffs with encrypted matching

ROLLOUT ROADMAP

Phase 1: Foundation (Q1 2025)

- 5% time-to-play reduction
- 10K beta testers

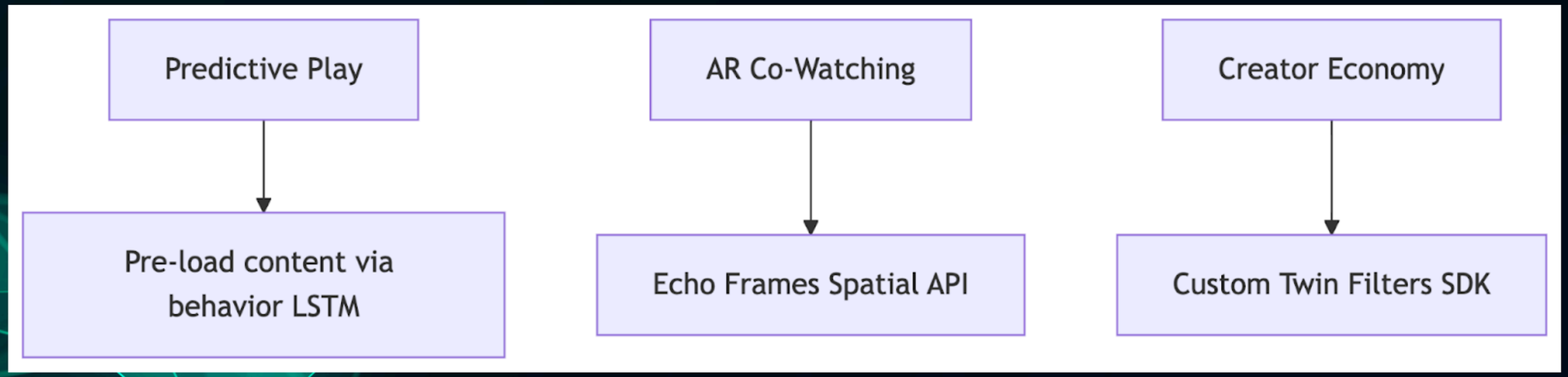
Phase 2: Scaling (Q2 2025)

- Watchlist Twin launch
- Global edge caching rollout
- Tech Milestone: 500K QPS matching

Phase 3: Ecosystem (Q3 2025)

- Alexa/Echo Show integration
- Developer SDK launch
- Tech Milestone: Cross-device RL agent

FUTURE VISION



2026+ Architecture Evolution

Tech Frontiers:

- Quantum: ANN on Braket (prototyping)
- HPC: Real-time ray tracing for UI
- Bio-API: Halo v3 stress correlation

RISK MITIGATION AND COMPILANCE

Technical Safeguards

Risk	Solution	AWS Service
Data Leakage	Homomorphic encryption + KMS	AWS Nitro Enclaves
LLM Hallucination	Fine-tuned RLHF + human eval	SageMaker Clarify
GDPR Compliance	Zero-knowledge proofs	OpenFHE + IAM
Latency Spikes	Auto-scaling + circuit breakers	CloudWatch + Hystrix
Cost Overruns	Spot fleet + MoE sparsity	EC2 Auto Scaling

Disaster Recovery

- *Multi-region DynamoDB*
- *ANN index snapshotting to S3 Glacier*

THANK YOU!

FOR YOUR ATTENTION