Lilo Engine

Production-Grade Therapeutic AI Platform

Multi-Agent AI for Elderly Mental Health in Senior Living Facilities

[**65K+** Lines of Code] [**15** Microservices] [**100%** Crisis Recall] [**<1s** Response]

Platform At-A-Glance — Key Metrics

CODEBASE
65,000+
lines

SERVICES
15
microservices

CRISIS RECALL
100%
zero false neg

RESPONSE TIME
<1 second
30x regulatory

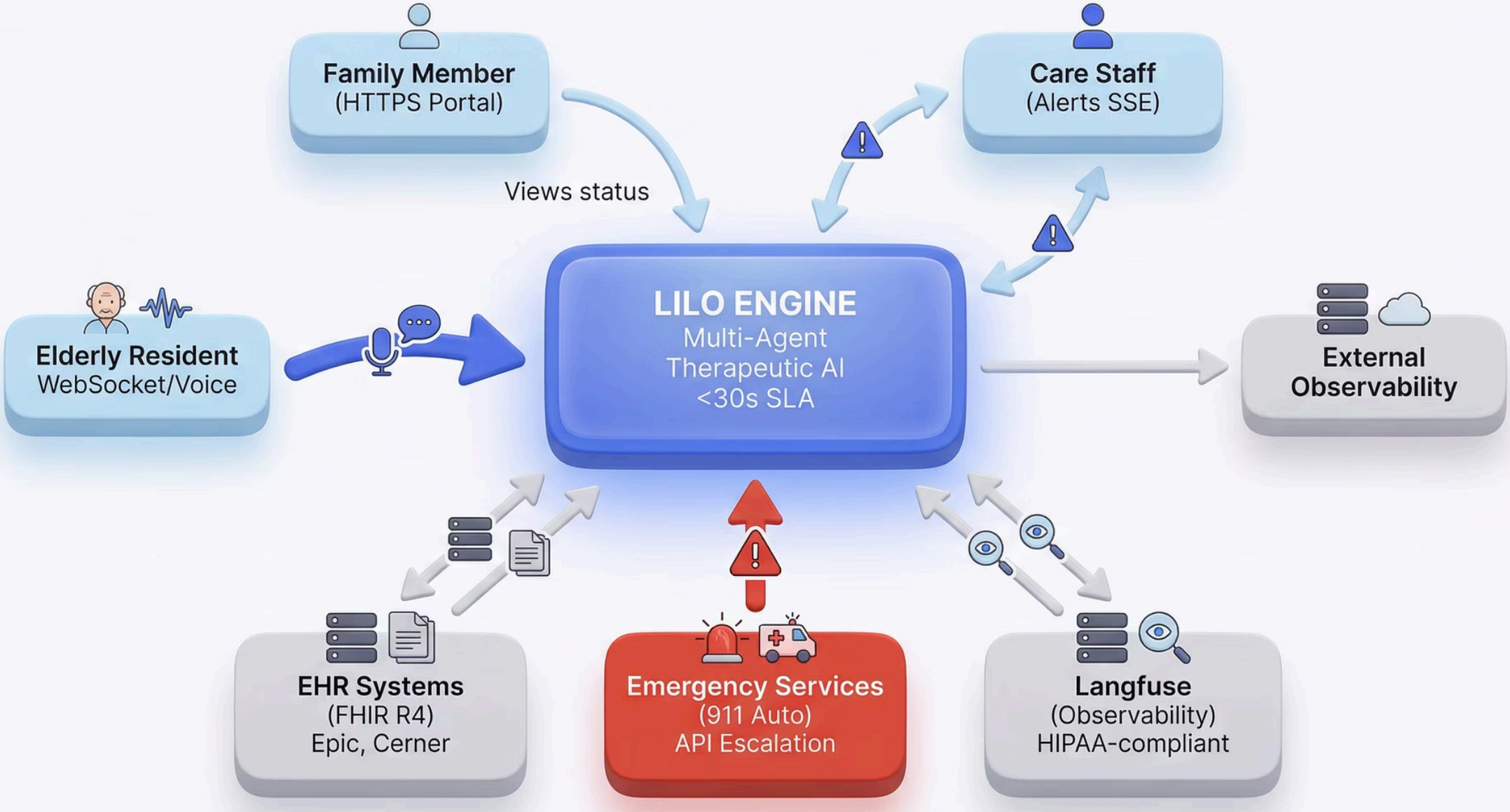
INTENT ACCURACY
92-95%
214 prototypes

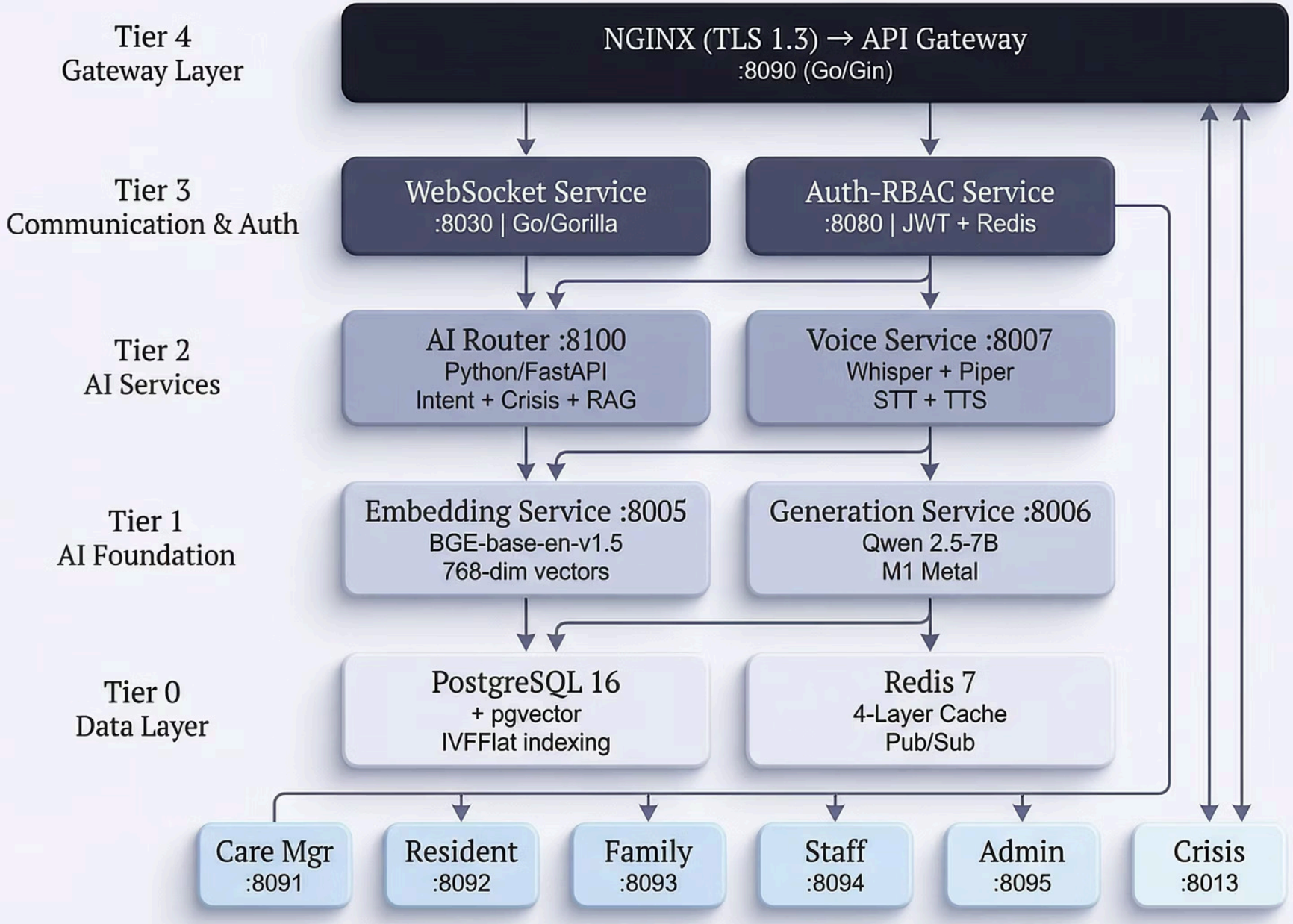
THERAPEUTIC AGENTS
7
evidence-based

CACHE HIT RATE
60-80%
<5ms on hit

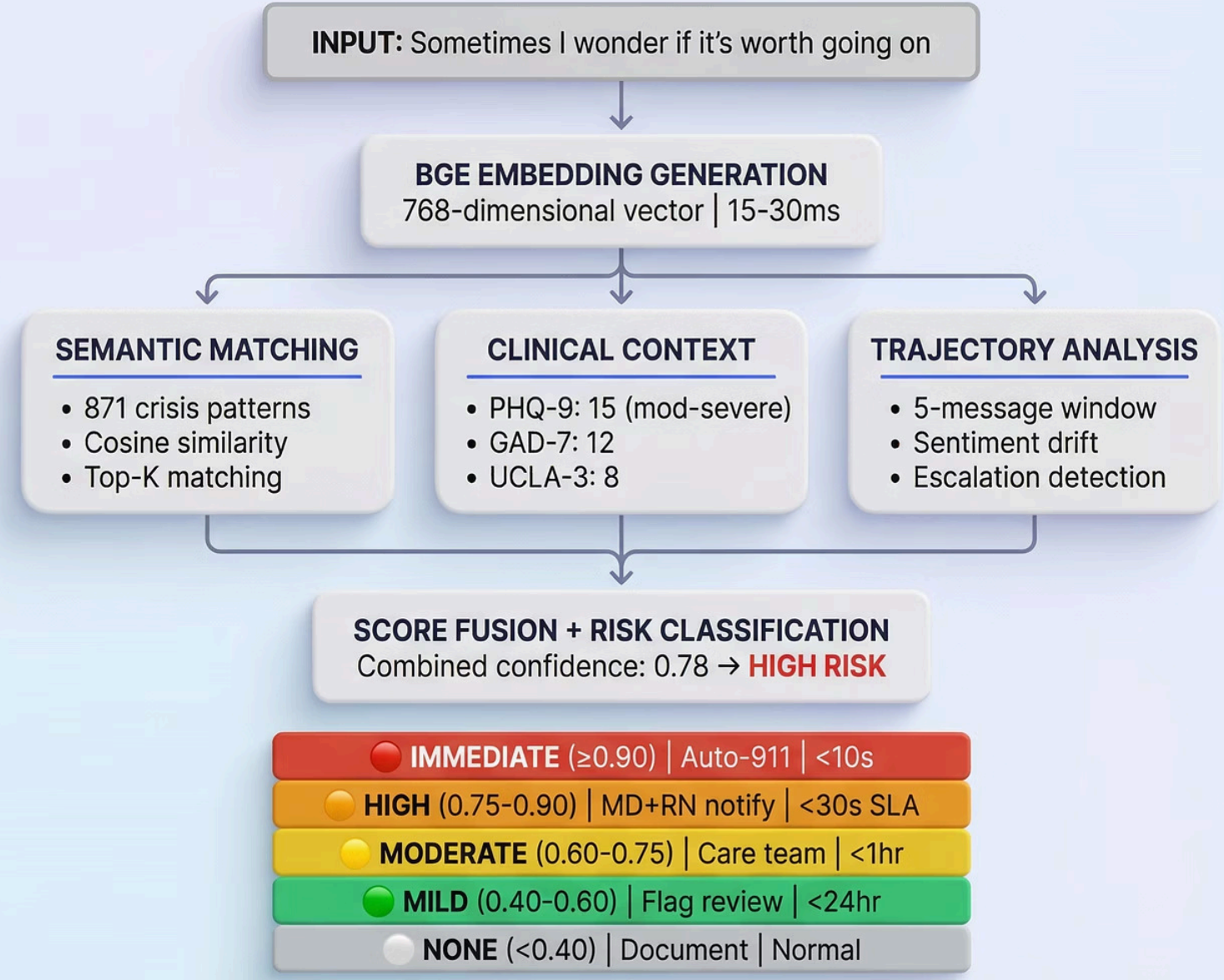
GENERATION LATENCY
~7.6s
streaming

[Python 3.12] [Go 1.25] [PyTorch 2.8] [PostgreSQL 16 + pgvector] Redis 7] Docker Qwen 2.5-7B] BGE Embeddings

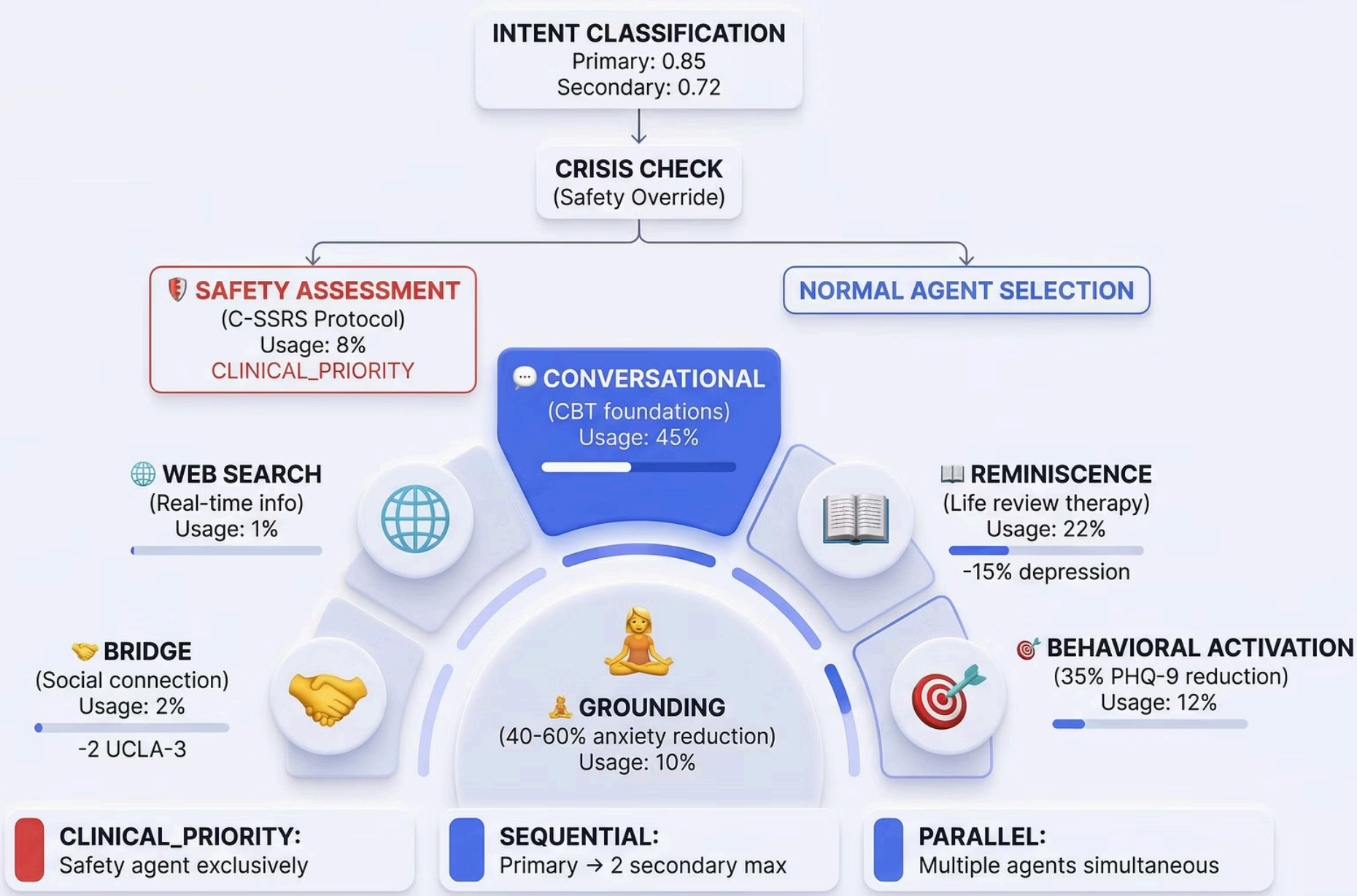

Family Member
(HTTPS Portal)
Views status
Care Staff
(Alerts SSE)
Elderly Resident
WebSocket/Voice
LILO ENGINE
Multi-Agent
Therapeutic AI
<30s SLA
External
Observability
EHR Systems
(FHIR R4)
Epic, Cerner
Emergency Services
(911 Auto)
API Escalation
Langfuse
(Observability)
HIPAA-compliant

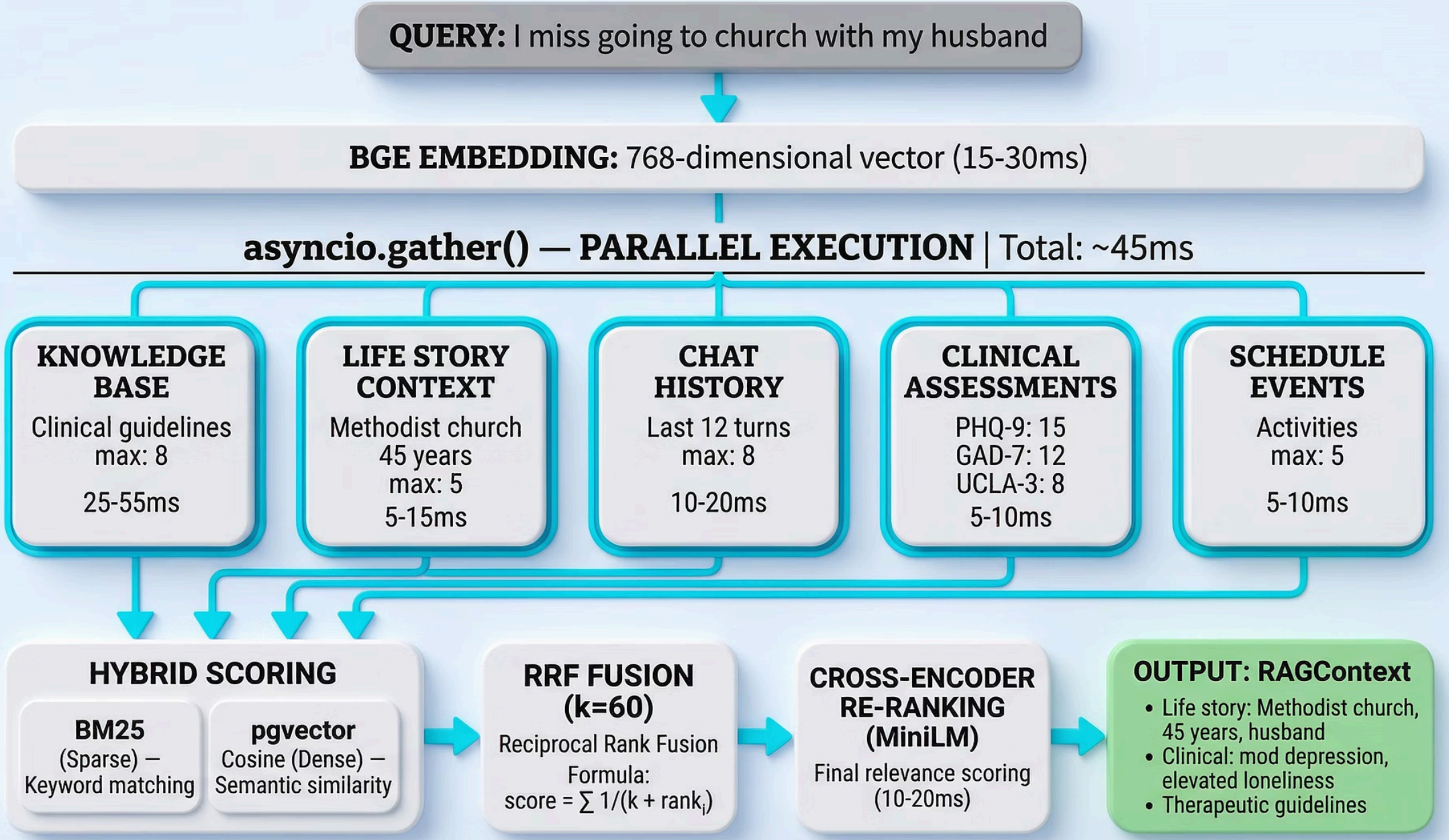

Tier 4
Gateway Layer
NGINX (TLS 1.3) → API Gateway
:8090 (Go/Gin)
Tier 3
Communication & Auth
WebSocket Service
:8030 | Go/Gorilla
Auth-RBAC Service
:8080 | JWT + Redis
Tier 2
AI Services
AI Router :8100
Python/FastAPI
Intent + Crisis + RAG
Voice Service :8007
Whisper + Piper
STT + TTS
Tier 1
AI Foundation
Embedding Service :8005
BGE-base-en-v1.5
768-dim vectors
Generation Service :8006
Qwen 2.5-7B
M1 Metal
Tier 0
Data Layer
PostgreSQL 16
+ pgvector
IVFFlat indexing
Redis 7
4-Layer Cache
Pub/Sub
Care Mgr
:8091
Resident
:8092
Family
:8093
Staff
:8094
Admin
:8095
Crisis
:8013

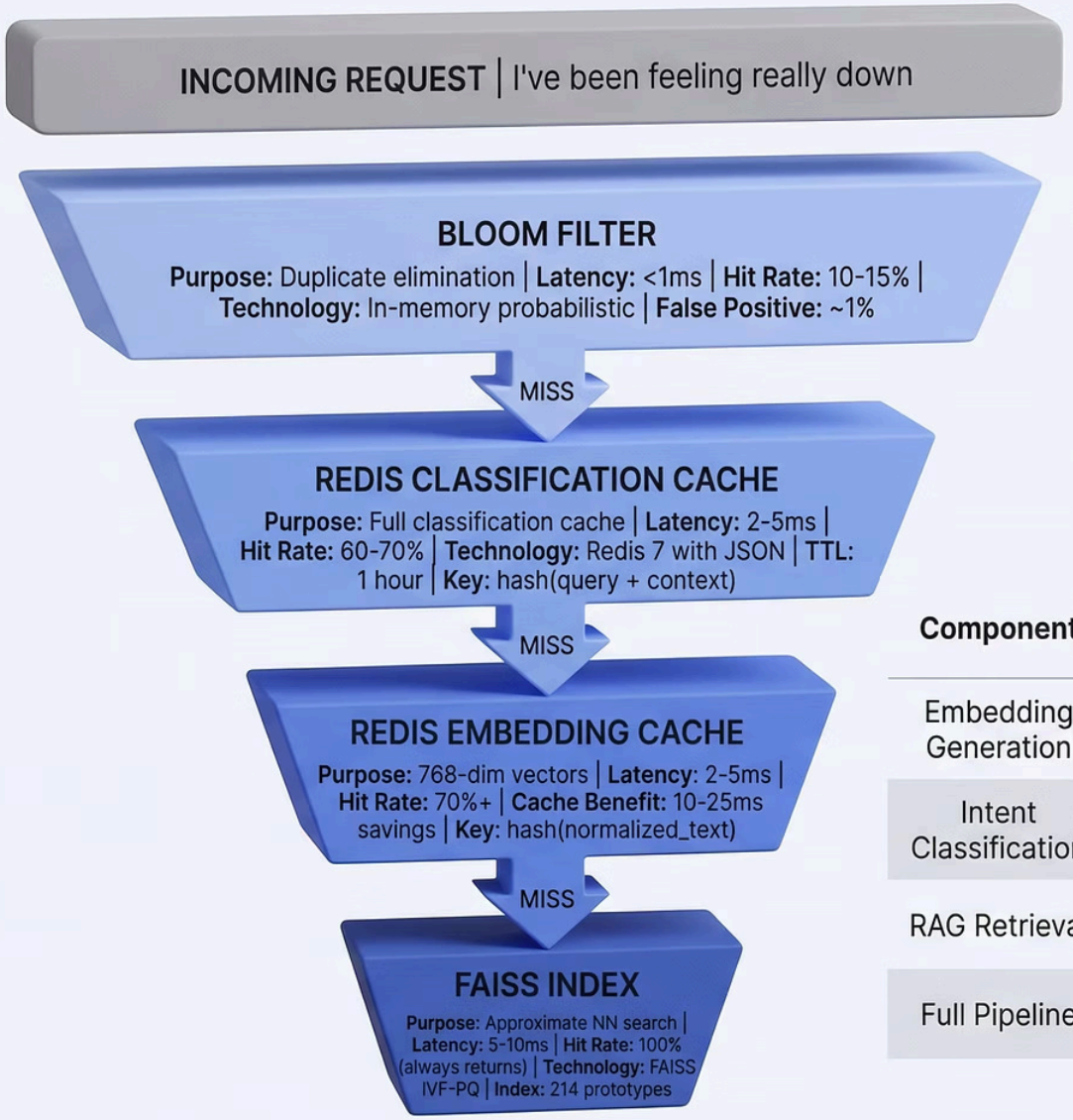

USER MESSAGE: I've been feeling really down lately
1. CLIENT CONNECTION (50ms) | WebSocket + JWT validation
50ms
2. API GATEWAY + HIPAA (5ms) | PHI redaction
5ms
3. SAFETY FIRST — CRISIS DETECTION (20-50ms)
• 871 crisis patterns | • Clinical context
• 5-message trajectory | • IF CRISIS → Parallel escalation
20-50ms
asyncio.gather — PARALLEL
15-30ms
4. EMBEDDING (15-30ms)
5. INTENT CLASSIFICATION (10-50ms)
10-50ms
6. AGENT ORCHESTRATION (5ms) | Select primary + secondary agents
5ms
7. RAG RETRIEVAL (30-55ms) | 5 PARALLEL STREAMS
Knowledge
Life Story
Chat History
Assessments
Schedule
30-55ms
8. PROMPT CONSTRUCTION (5-15ms)
5-15ms
9. LLM GENERATION (~7.6s) | Qwen 2.5-7B | Metal GPU
~7.6s
10. POST-PROCESSING (<5ms) | Safety validation
<5ms
11. DELIVERY + AUDIT (10ms) | WebSocket/SSE | Langfuse
10ms
RESPONSE: I hear you, and I'm glad you shared...



Lilo Engine	Industry Standard
Recall: **100%**	60-70%
Precision: **95%+**	70-80%
Response Time: **<1s**	15-30min
False Positives: **<5%**	20-30%


INTENT CLASSIFICATION
Primary: 0.85
Secondary: 0.72
CRISIS CHECK
(Safety Override)
SAFETY ASSESSMENT
(C-SSRS Protocol)
Usage: 8%
CLINICAL_PRIORITY
NORMAL AGENT SELECTION
CONVERSATIONAL
(CBT foundations)
Usage: 45%
WEB SEARCH
(Real-time info)
Usage: 1%
REMINISCENCE
(Life review therapy)
Usage: 22%
-15% depression
BRIDGE
(Social connection)
Usage: 2%
-2 UCLA-3
GROUNDING
(40-60% anxiety reduction)
Usage: 10%
BEHAVIORAL ACTIVATION
(35% PHQ-9 reduction)
Usage: 12%
CLINICAL_PRIORITY:
Safety agent exclusively
SEQUENTIAL:
Primary → 2 secondary max
PARALLEL:
Multiple agents simultaneous


QUERY: I miss going to church with my husband
BGE EMBEDDING: 768-dimensional vector (15-30ms)
asyncio.gather() — PARALLEL EXECUTION | Total: ~45ms
KNOWLEDGE BASE
Clinical guidelines
max: 8
25-55ms
LIFE STORY CONTEXT
Methodist church
45 years
max: 5
5-15ms
CHAT HISTORY
Last 12 turns
max: 8
10-20ms
CLINICAL ASSESSMENTS
PHQ-9: 15
GAD-7: 12
UCLA-3: 8
5-10ms
SCHEDULE EVENTS
Activities
max: 5
5-10ms
HYBRID SCORING
BM25
(Sparse) — Keyword matching
pgvector
Cosine (Dense) — Semantic similarity
RRF FUSION (k=60)
Reciprocal Rank Fusion
Formula:
score = ∑ 1/(k + rank_i)
CROSS-ENCODER RE-RANKING (MiniLM)
Final relevance scoring
(10-20ms)
OUTPUT: RAGContext
• Life story: Methodist church, 45 years, husband
• Clinical: mod depression, elevated loneliness
• Therapeutic guidelines

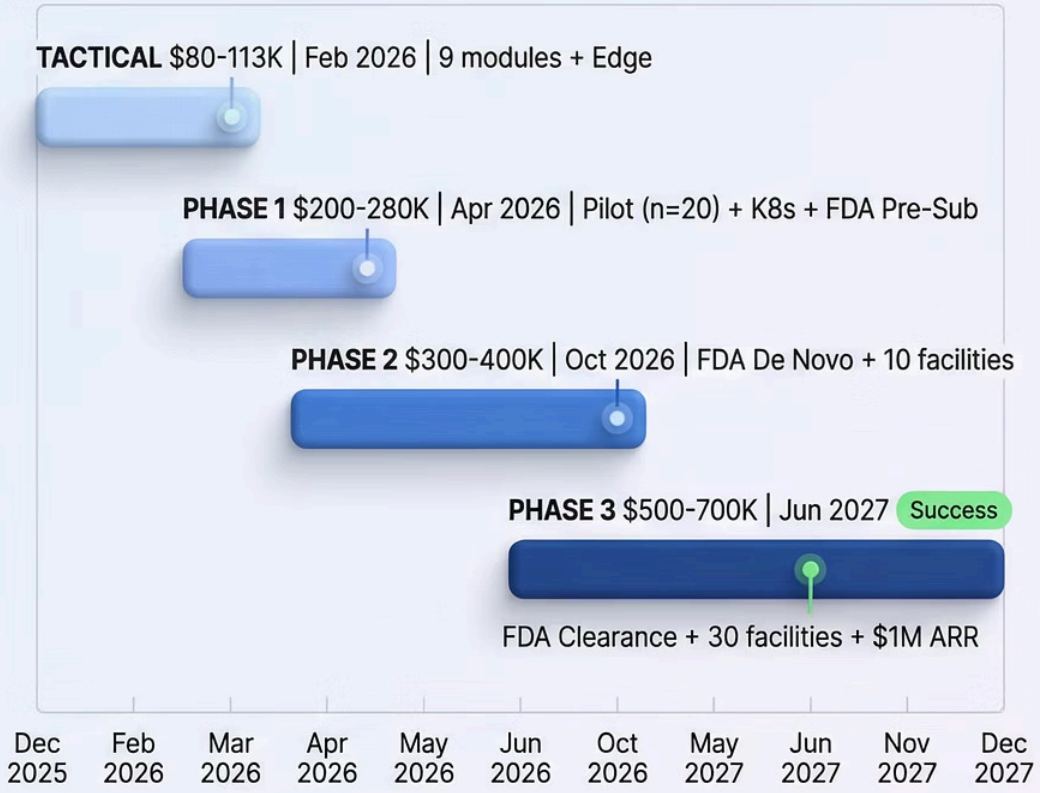


Component	Without Cache	With Cache	Savings
Embedding Generation	15-30ms	2-5ms	10-25ms
Intent Classification	40-50ms	10-15ms	30-40ms
RAG Retrieval	60-70ms	30-45ms	15-25ms
Full Pipeline	150-200ms	80-120ms	~50%

Technical Differentiators — Why Lilo Engine Wins

DIFFERENTIATOR	LILO ENGINE	INDUSTRY STANDARD
Crisis Detection Recall	100% Zero false negatives	60-70% Missed crises common
Response Time	<1s 30x regulatory	15-30 minutes Manual review
False Positive Rate	<5% Low alert fatigue	20-30% Staff overwhelmed
HIPAA Compliance	Built-in Day 1	Often retrofitted
Edge Deployment	8GB 90/10 split	N/A Cloud-only
Therapeutic Agents	7 Evidence-based	1 generic chatbot
Clinical Integration	PHQ-9 GAD-7 C-SSRS	Limited
RAG Personalization	5 parallel streams	Generic only

Strategic Roadmap



$1.1-1.5M to FDA Clearance