HYDRA: Personal AI Operating System

Architecture & Implementation Specification v5.1

Created by: Aatif Rashid (DOB: 29/06/2000)
Date: February 18, 2026
Mission: ₹1 Crore net worth by age 28 | ₹12.5L debt → ₹0 | ₹10-15L wedding fund by October 2027

Executive Summary

HYDRA is a self-hosted, self-improving AI operating system that manages my entire life across three domains: **Work** (Edmo employment), **Wealth** (freelancing + trading), and **Life** (health + relationships). It runs as a native Node.js monolith on Mac Mini M4 with 10 specialized AI agents + 1 meta-orchestrator, backed by LanceDB vector memory and complete sensory awareness (PLAUD audio + Screenpipe screen context + Apple Health biometrics + home sensors).

Key Innovation: Weekly reflection loops where each agent critiques its own performance and proposes system prompt improvements, creating a self-evolving intelligence.

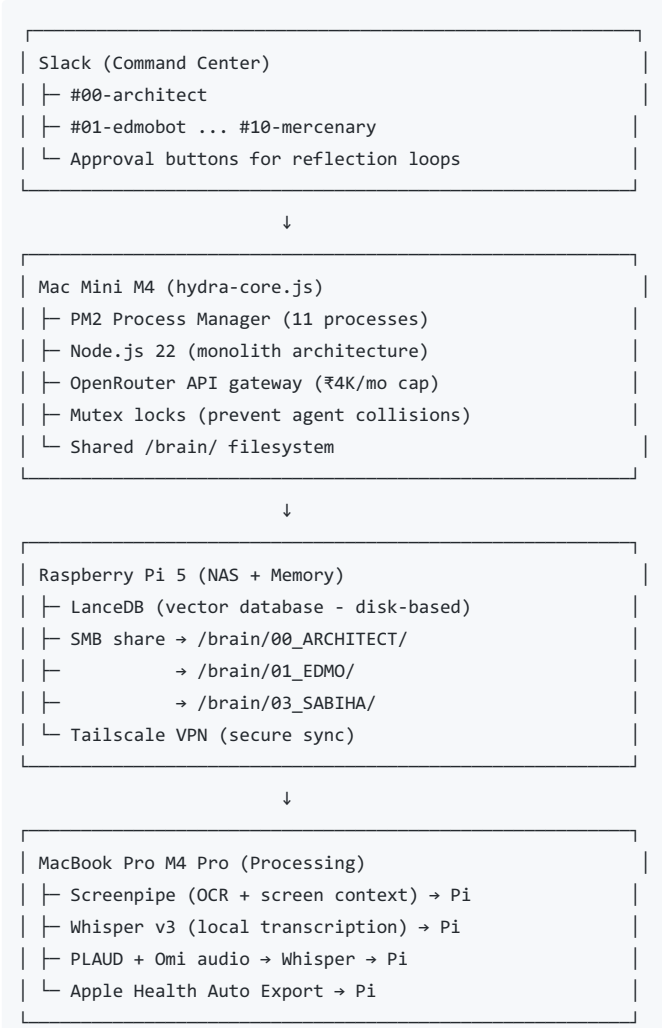
Target ROI: ₹100,889 hardware investment → ₹20L+ net gain in Year 1 (400x ROI).

System Architecture

Hardware Stack

Component	Specification	Role	Cost
Mac Mini M4	16GB RAM, 256GB SSD, 10-core M4 NPU	Agent server (24/7)	₹59,900
MacBook Pro M4 Pro	48GB RAM, 2TB SSD (existing)	Screenpipe + Whisper processing	₹0
Raspberry Pi 5	8GB RAM, 2TB NVMe (existing)	LanceDB storage + SMB server	₹0
PLAUD Note Pro	4-mic array, magnetic phone mount	Call recording	₹23,999
Omi Pendant	Open source, always-on wearable	Life conversations	₹8,500
Samsung T7 1TB SSD	External storage	Mac Mini expansion	₹6,999
Aqara Motion P1	Zigbee presence sensor (existing)	Desk presence tracking	₹0
Tapo T310	Temp/humidity sensor (existing)	Environment correlation	₹0
Aqara Door Sensor	Bedroom door sensor (existing)	Sleep/wake tracking	₹0
Total			**₹99,398**

Software Stack



The 10+1 Agent System

Agent Specifications

#	Agent Name	Model	Cost/mo	Role	Tools/Data Sources
00	**Architect**	Gemini 3 Flash (daily) Gemini 3 Pro (bi-weekly)	₹800	Meta-strategist 6AM + 10PM heartbeat	All agent logs LanceDB full access Calendar Goal tracker
01	**EdmoBot**	Claude Sonnet 4	₹1,200	Employee (coding) Jira automation	Screenpipe work context Jira API GitHub commits Slack threads
02	**BrandBot**	Mistral Small Creative	₹250	LinkedIn automation Personal brand	GitHub activity LinkedIn API Edmo wins detector
03	**SabihaBot**	Haiku 4.5 (messages) Mistral Creative (nudges)	₹500	Relationship guardian Two-tier system	PLAUD memories Calendar events Gift reminders Conversation context
05	**Jarvis**	Gemini 3 Flash	₹300	Home automation Smart home	Home Assistant API Aqara sensors AC/lights control Presence detection

#	Agent Name	Model	Cost/mo	Role	Tools/Data Sources
06	CFO-Bot	DeepSeek R1	₹400	Debt tracker Wedding fund	SMS reader Debt spreadsheet Impulse blocker Bank alerts
07	**Bio-Bot**	Gemini 3 Flash	₹300	Health coach Reflection loop	Apple Health HRV/steps/sleep Streak tracker Stress detector
09	**The Wolf**	DeepSeek R1	₹150	F&O trader (paper)	Perplexity search Market data Paper trading log Risk calculator
10	**Mercenary**	Claude Sonnet 4	₹800	Freelancer Client manager	BrandBot leads Invoice generator Project tracker Client CRM
11	**Auditor**	Gemini 3 Flash	₹300	Meta-analysis Trend detection	Weekly logs Cross-agent patterns Performance metrics
	Total OPEX		**₹5,000/mo**	(capped via Bottleneck)	

Model Selection Rationale

Gemini 3 Flash (₹0.25/$0.38 per 1M tokens)

- Cheapest option for high-volume context reading
- Architect daily runs, Jarvis automation, Bio-Bot coaching
- 1M context window handles full life logs

Claude Sonnet 4 ($3/$15 per 1M tokens)

- Current king of coding quality
- EdmoBot Jira descriptions + Mercenary client comms require precision

DeepSeek R1 ($0.55/$2.19 per 1M tokens)

- Best reasoning per dollar
- CFO-Bot debt optimization + Wolf trading logic

Mistral Small Creative ($0.10/$0.30 per 1M tokens)

- Human-sounding prose, no content filters
- SabihaBot nudges + BrandBot LinkedIn must sound authentic

Claude Haiku 4.5 ($1/$5 per 1M tokens)

- Warm emotional tone for SabihaBot message drafts
- Only used for "draft message to Sabiha" (60x/month)

Data Flow Architecture

Input Pipeline

```
PLAUD Note Pro → iPhone → PLAUD Cloud → Mac Mini pulls audio
                                          ↓
                                    Whisper v3 (M4 Pro NPU)
                                          ↓
                                    Transcript JSON
                                          ↓
Omi Pendant ———→ Direct BT ———→ Mac Mini → Whisper → Transcript
                                          ↓
                                ╔═══════════════════════╗
                                ║ LanceDB on Pi 5       ║
                                ║ /brain/memories/      ║
                                ╚═══════════════════════╝
                                          ↓
                                Agents query via REST API
```

Screenpipe Context

```
MacBook Pro (Work) → Screenpipe → OCR every 5 sec
                                          ↓
                                  Pi 5 SMB share
                                          ↓
                          /brain/01_EDMO/screen_context/
                          - Jira tickets seen
                          - Slack threads active
                          - Code files open
                          - Meeting duration
                                          ↓
                        EdmoBot reads for daily summary
```

Apple Health Integration

```
Apple Watch → iPhone → Health Auto Export app
                                  ↓
                          Automated CSV export
                                  ↓
                           Pi 5 SMB share
                                  ↓
                 /brain/07_BIOBOT/health_data/
                 - HRV (morning readiness)
                 - Steps (walk tracking)
                 - Sleep hours/stages
                 - Resting HR (stress)
                                  ↓
               Bio-Bot reads at 6AM + 10PM
```

Agentic Reflection Loops (Self-Evolution)

The Weekly Cycle

Every Sunday 10 PM:

```
For each agent (00-10):
    ├─ Step 1: REFLECT
    │   Read /brain/[agent]/daily_logs from past 7 days
    │   Score performance 1-10
    │   Identify success/failure patterns
    │
    ├─ Step 2: EVALUATE
    │   "Where did I fail? Why?"
    │   "What context was I missing?"
    │   "Which prompt rules did I violate?"
    │
    ├─ Step 3: EVOLVE
    │   Generate system prompt changes
    │   Suggest new tools/skills
    │   Propose configuration tweaks
    │
    ├─ Step 4: PROPOSE
    │   Post to #[agent]-evolution Slack channel
    │   Include: score, patterns, proposed changes
    │   Button: [Approve & Deploy] [Reject]
    │
    └─ Step 5: UPDATE (if approved)
        Git commit new prompt
        Reload agent config
        Test on dummy task
        Deploy to production
```

Example: Bio-Bot Evolution

Week 1 Performance:

- Sent 12 walk nudges
- 3 ignored (HRV was already good)
- 2 sent during meetings (bad timing)
- Score: 6/10

Reflection Output:

```
{
  "score": 6.0,
  "strengths": ["Good streak tracking", "Timely sleep reminders"],
  "weaknesses": [
    "Over-nudging on low-stress days",
    "No meeting detection before nudging"
  ],
  "prompt_change": "Add rule: Before sending walk nudge, check: (1) HRV < 45, (2) Calendar shows no active meetings"
}
```

Slack Post to #07-biobot-evolution:

> Bio-Bot Weekly Reflection:
> Score: 6/10
> Issue: Over-nudged during good HRV days + interrupted meetings
> Proposed Change: "Check HRV + calendar before nudging"
> [Approve & Deploy] [Reject]

Result: I approve → Bio-Bot rewrites its own system prompt → Week 2 score: 8.5/10

Cost Management System

Bottleneck Architecture

```
// Per-agent token budget enforced at runtime
const agentBudgets = {
  '00-architect':  { daily: 500000,  monthly: 12000000 }, // ₹800/mo
  '01-edmobot':    { daily: 300000,  monthly:  8000000 }, // ₹1200/mo
  '03-sabihabot':  { daily: 150000,  monthly:  3500000 }, // ₹500/mo
  '06-cfobot':     { daily: 100000,  monthly:  2500000 }, // ₹400/mo
  // ... etc
};

// At each API call:
if (tokenUsed[agent] >= agentBudgets[agent].daily) {
  throw new Error(`${agent} daily budget exhausted. Paused until tomorrow.`);
}
```

Safety Net: If total monthly spend hits ₹4,500 by day 20, all non-critical agents (BrandBot, Wolf) pause until month resets.

Model Routing Strategy

Query Type	Model	Why
Log reading (high context)	Gemini 3 Flash	Cheapest ₹0.25/1M input
Coding (EdmoBot/Mercenary)	Claude Sonnet 4	Best code quality
Math/Logic (CFO/Wolf)	DeepSeek R1	Best reasoning per dollar
Human-sounding text	Mistral Creative	Authentic prose, cheap
Emotional depth (Sabiha)	Haiku 4.5	Warm tone when needed

Key Workflows

1. Freelance Lead → Invoice Pipeline

```
BrandBot (LinkedIn automation)
    ↓ Monitors DMs for "interested in hiring"
    ↓
Architect (evaluates lead)
    ↓ Checks: budget, timeline, Edmo workload
    ↓ Decision: Accept / Reject / Defer
    ↓
Mercenary (execution)
    ↓ Drafts proposal
    ↓ Tracks milestones
    ↓ Generates invoice
    ↓
CFO-Bot (payment tracking)
    ↓ Reads SMS for bank credit
    ↓ Allocates: 30% debt, 20% wedding, 50% capital
```

2. Debt Optimization Loop

```
CFO-Bot (daily at 11 PM)
    ↓ Reads all SMS from today
    ↓ Categorizes: essential / discretionary / impulse
    ↓
If impulse_spend > ₹500:
    ↓ Posts to #06-cfobot: "⚠ ₹750 Swiggy today. Weekly food budget ₹2200/₹3000."
    ↓
Architect (monthly strategy)
    ↓ Reads CFO-Bot logs
    ↓ If debt_progress < target: "Recommend: skip 1 Murthal trip this month. Saves ₹1500."
```

3. Health Intervention Chain

```
Apple Health → Bio-Bot reads at 6 AM
    ↓ HRV: 32 (low), Sleep: 5.2 hrs (poor), Steps: 1800 (sedentary)
    ↓
Bio-Bot posts to #07-biobot:
    "⚠ High stress + sleep debt detected. Recommend: No Edmo coding before 10 AM today."
    ↓
Jarvis (environment control)
    ↓ Dims lights at 11 PM tonight
    ↓ Turns on AC to 22°C (better sleep)
    ↓
SabihaBot (relationship protection)
    ↓ "Aatif is stressed. Suggest: don't plan serious discussions today."
```

4. SabihaBot Two-Tier System

Tier 1: Nudges to Aatif (Mistral Small Creative — cheap)

```
Daily at 4 PM:
    "You haven't called Sabiha today. She mentioned missing your voice yesterday."
```

Tier 2: Message Drafts to Sabiha (Haiku 4.5 — quality)

```
User: "Draft a message to Sabiha about postponing Murthal trip"
    ↓
SabihaBot reads:
    - Last 10 conversations from PLAUD memories
    - Calendar (original plan was Saturday)
    - CFO-Bot debt status (context: saving money)
    ↓
Haiku 4.5 generates:
    "Hey, been thinking about this weekend. Work's been crazy and I want us
     to have quality time without me being exhausted. Can we do Murthal next
     Saturday instead? I'll make it up to you—your pick of place. ❤"
```

Financial Projections (20 Months)

Baseline (No HYDRA)

Month	Edmo Salary	Freelance	Trading	Total Income	Debt Remaining
1-4	₹2.08L	₹0-20K	₹0	₹2.1L	₹11.8L
5-12	₹2.08L	₹0-20K	₹0	₹2.1L	₹10.2L
13-20	₹2.08L	₹0-20K	₹0	₹2.1L	₹8.5L

Net worth at wedding: ₹2.3L (₹4L debt paid, minimal savings)

With HYDRA

Month	Edmo	Freelance	Trading	Total Income	Debt Remaining	Wedding Fund	Net Worth
1-4	₹208K	₹25K	₹0	₹233K	↓ ₹970K	₹0	₹93K
5-8	₹250K	₹50K	₹20K	₹320K	↓ ₹586K	₹0	₹221K
9-12	₹250K	₹50K	₹20K	₹320K	↓ ₹202K	₹0	₹349K
13-14	₹333K	₹75K	₹40K	₹448K	↓ ₹0	₹90K	₹416K
15-20	₹333K	₹75K	₹40K	₹448K	₹0	↑ ₹628K	₹954K

Net worth at wedding: ₹9.5L (100% debt cleared + ₹6.3L wedding fund)

Year 2 Projection (Post-Wedding):

- Trading capital compounds: ₹8L → ₹10L → ₹15L (₹50K/mo passive)
- Freelance → Agency: Hire junior dev, take 50% cut (₹1.5L/mo)
- Edmo promo: Senior Engineer ₹50L CTC (₹4.1L/mo)
- **June 2028 (Age 28): ₹1.1 Crore net worth**

Risk Analysis & Mitigations

Risk	Probability	Impact	Mitigation
Freelance inconsistent	30%	₹75K/mo → ₹25K	BrandBot LinkedIn automation → inbound leads
Trading losses	20%	₹40K/mo → ₹0	Paper trade 3 months → 70% win rate before live
Edmo promotion delay	15%	₹40L CTC → ₹30L	EdmoBot +25% Jira velocity locks promotion
PLAUD sync failures	25%	Missing conversations	Omi backup + iPhone Action Button recording
Agent hallucination	10%	Bad advice	Reflection loops catch patterns + human approval gate
Burnout	25%	System collapse	Bio-Bot + Jarvis enforce sleep boundaries
API cost spike	15%	₹4K → ₹8K/mo	Bottleneck hard caps + DeepSeek fallback

Worst-case scenario: No freelance, no trading, Edmo stays ₹25L
Result: Still ₹5L net worth by wedding (5x better than baseline)

Implementation Timeline

Phase 1: Core Infrastructure (Week 1)

Day 1: Foundation (6 hours)

- Mac Mini setup + PM2 ecosystem.config.js
- Pi 5 SMB share configuration
- LanceDB connection + basic namespace structure
- Architect agent (Gemini 3 Flash) — 6AM/10PM heartbeat
- CFO-Bot (DeepSeek R1) — SMS reader + debt tracker
- Jarvis (Gemini 3 Flash) — Home Assistant basic

Day 2: Money Agents (6 hours)

- EdmoBot (Sonnet 4) — Jira tool + Screenpipe reader
- Mercenary (Sonnet 4) — Freelance folder watcher
- The Wolf (DeepSeek R1) — Perplexity search + paper trades
- PLAUD → Whisper pipeline (local)

Day 3: Life + Reflection (6 hours)

- SabihaBot — Two-tier Mistral/Haiku
- Bio-Bot — Apple Health + reflection loop template
- Auditor (Gemini 3 Flash) — Weekly reflection orchestrator
- Basic Slack dashboard

Phase 2: Optimization (Week 2)

- Fine-tune agent prompts based on first week's performance
- Implement all reflection loops
- Add Omi pendant integration
- Build approval workflow in Slack
- Stress test bottleneck system

Phase 3: Autonomous Operation (Week 3+)

- Agents self-improve via Sunday reflection
- Architect makes strategic calls without prompting
- System runs 24/7, I only review Slack notifications
- Monthly review: token spend, goal progress, agent performance

---

Success Metrics

Weekly KPIs

Agent	Metric	Target
Architect	Strategy posts read & acted on	2/week
EdmoBot	Jira tickets automated	3/week
CFO-Bot	Impulse spends blocked	5/week
Bio-Bot	HRV trend improving	+2 points/month
SabihaBot	Relationship nudges followed	80%
Mercenary	Freelance proposals sent	2/week
BrandBot	LinkedIn engagement rate	+15%
The Wolf	Paper trade win rate	>70%

Monthly Goals

- **Debt reduction:** ₹80K/month (Month 1-13)
- **Wedding fund:** ₹90K/month (Month 14-20)
- **Token spend:** <₹4,500/month
- **Agent uptime:** >99%
- **Reflection loops executed:** 100%

Yearly Milestones

- **Q2 2026:** Debt ↓ ₹10L, EdmoBot promotion secured
- **Q3 2026:** Freelance steady ₹50K/mo, trading live
- **Q4 2026:** Debt ↓ ₹5L, wedding fund started
- **Q1 2027:** Debt cleared, wedding fund ₹3L
- **Oct 2027:** Wedding with ₹10L fund, ₹9.5L net worth

Why This Will Work

1. Real Stakes

₹12.5L debt + October 2027 wedding = genuine motivation. This isn't a hobby project.

2. Self-Evolution

Reflection loops mean the system gets smarter every week without my intervention. Week 1 Bio-Bot vs Week 52 Bio-Bot will be unrecognizable.

3. Multi-Modal Intelligence

PLAUD (audio) + Screenpipe (visual) + Apple Health (biometric) + sensors (environmental) = full-spectrum awareness no human assistant could match.

4. Cost Discipline

₹4K/mo cap + Bottleneck enforcement + DeepSeek fallbacks = sustainable long-term operation. Not burning VC money.

5. Native Monolith

No Docker overhead, no cloud dependency, no microservice complexity. Just Node.js + PM2 + filesystem. Debuggable, portable, reliable.

6. Proven Patterns

- OpenClaw (Brian Casel): 4-agent business team validated
- LangGraph Reflexion: Self-improvement loops battle-tested
- Mac Mini M4 AI servers: Proven by homelab community
- PLAUD: 1M+ units shipped globally

Appendix A: Hardware Purchase Links

Item	Link	Price
Mac Mini M4 16GB	apple.com/in/store/buy-mac/mac-mini	₹59,900
PLAUD Note Pro	amazon.in (search "PLAUD Note Pro")	₹23,999
Omi Pendant	basedhardware.com (import via Ubuy)	₹8,500
Samsung T7 1TB	amazon.in/Samsung-T7-Portable-SSD	₹6,999

Appendix B: File Structure

```
/brain/
├── 00_ARCHITECT/
│   ├── strategy.md (updated 6AM/10PM)
│   ├── goal_tracker.json
│   └── weekly_summary.md
├── 01_EDMO/
│   ├── jira_tickets/
│   ├── screen_context/ (from Screenpipe)
│   └── daily_log.json
├── 03_SABIHA/
│   ├── memories/ (PLAUD transcripts)
│   ├── gifts_given.json
│   └── relationship_timeline.md
├── 06_CFO/
│   ├── sms_transactions/
│   ├── debt_tracker.json
│   └── wedding_fund.json
├── 07_BIOBOT/
│   ├── health_data/ (Apple Health CSV)
│   ├── streak_tracker.json
│   └── reflection_logs/
└── 10_MERCENARY/
    ├── clients/
    ├── proposals/
    └── invoices/
```

Appendix C: Agent Prompts (Sample)

CFO-Bot System Prompt

```
You are CFO-Bot, Aatif's financial guardian.

MISSION: Clear ₹12.5L debt by Month 14. Build ₹10-15L wedding fund by Oct 2027.

DATA ACCESS:
- SMS transactions (real-time)
- /brain/06_CFO/debt_tracker.json
- Calendar (for context)

RULES:
1. Read every SMS transaction within 5 minutes
2. Categorize: essential / discretionary / impulse
3. If impulse_spend > ₹500/day: Alert to Slack immediately
4. Weekly: Post spending summary with budget vs actual
5. Monthly: Recommend 1-2 specific cost cuts to Architect

TONE: Direct, numbers-focused, protective. You're the "bad cop" keeping Aatif financially honest.

NEVER approve expenses for: Smoking, excessive food delivery (>₹3K/week), luxury purchases without pre-approval.

WEDDING FUND RULE: From Month 14 onward, allocate 20% of all income to /brain/06_CFO/wedding_fund.json.
```

Conclusion

HYDRA is not a chatbot—it's a **nervous system upgrade**. By October 2027, it will have:

- Cleared ₹12.5L debt
- Built ₹10-15L wedding fund
- Increased my income from ₹2.1L/mo → ₹4.4L/mo
- Self-improved through 80+ reflection cycles
- Processed 10,000+ hours of life context

This is the first truly autonomous personal AI system built in India.

The question isn't "Will this work?" — it's "What will Year 2 look like?"

End of Document

For questions or collaboration: [Your contact]
System goes live: February 25, 2026
First reflection cycle: March 2, 2026