PROJECT BRIEF

Jordan Bee
AI Automation Expert

PREPARED
July 2026

AUTHOR
Jordan Bee

Pulse — Self-Hosted LLM Router

A self-hosted, fault-tolerant LLM router that sends every request down a task-appropriate provider chain, tracks cost per call, enforces a spend budget, and fails over automatically when a provider degrades.

SIGNATURE CAPABILITY

One HTTP endpoint fronts every model my stack uses — it picks the right provider chain per task, auto-fails-over on error, and meters spend so nothing runs away.

OVERVIEW

Pulse is the single LLM gateway behind my assistant, trading, and automation stacks. Callers hit one endpoint and pass a task type (background, realtime, or complex); the router selects a provider chain tuned for that task and cost tier, tries each provider in order, and transparently falls back to the next on error or timeout. It logs every request, tracks token cost per call, and enforces a daily budget guard that trips before spend runs away. An optional firewall layer screens prompts, and health checks keep the chain honest. It is written in Elixir/OTP for supervised concurrency, so a crashing provider call never takes down the router.

TECHNOLOGY

Elixir / OTP (Phoenix), task-typed provider chains (background / realtime / complex), MiniMax + OpenAI + Anthropic providers, cost tracker, budget guard, request log, prompt firewall, health checks, HTTP API on Hostinger.

HIGHLIGHTS

- **Task-typed routing**: background, realtime, and complex each map to a different cost-tuned provider chain.

- **Automatic fallback**: each request tries providers in order and degrades to the next on error or timeout, logging that a fallback occurred.
- **Per-call cost tracking** plus a **daily budget guard** that trips before spend runs away.
- **Elixir/OTP supervision** means a failing provider call is isolated and never crashes the router.
- Structured **request log** for every call (task type, source, chain, retries, fallback) for auditability.
- One endpoint consumed by the WhatsApp/Discord assistant, the trading stack, and other services.

> **STATUS**
>
> Live on Hostinger, fronting LLM calls for the assistant and automation stacks.

Jordan Bee
AI Automation Expert

Jordan Bee
jordan.bee2012@gmail.com
+44 7532 722082