

Marco Hergi

Applied AI Engineer | Forward-Deployed / Founding Engineer

New York, NY | (917) 993-3893 | hergienterprises@gmail.com

linkedin.com/in/marcohergi | github.com/MARCCHERGGI | marco-hergi.vercel.app

SUMMARY

Self-taught applied AI engineer shipping production LLM systems solo and end-to-end: autonomous agents with persistent memory and tool use, real-time voice pipelines, on-device inference, and MCP tooling. **29 public repositories and 12 live products shipped since March 2025**, including a paid, operated AI assistant for a founder client. I route problems across models for cost and reliability, harden for production, and deploy same day. Seeking a forward-deployed, founding, or applied AI engineering role, customer-facing, solo to small team.

EXPERIENCE

Independent Applied AI Engineer (self-employed) - New York, NY

March 2025 to Present

- Designed, built, and operate **12 live AI products** end to end, solo: product decisions, model selection, backend, frontend, deployment, and on-call hardening. 29 public repositories, shipping continuously.
- Forward-deployed client work:** built and continue to operate a production AI assistant for a paying founder (PWA plus WhatsApp and Telegram delivery, morning-brief cron, domain knowledge of his ventures). Anthropic tool-use loop, durable Postgres memory, SSE streaming. Revenue-generating since May 2026.
- Built production agent infrastructure: multi-round tool-calling loops, retrieval grounding, persistent per-user memory, multi-provider routing with automatic failover, rate limiting, circuit breakers, and HMAC-signed usage limits across Groq, Anthropic, OpenAI, DeepSeek, and ElevenLabs.
- Authored and published **MCP servers** for the agent-interoperability standard, including a consensus-routing server that orchestrates 11 models with best-of-N routing and a 0 to 100 scoring engine.
- Build in public to **14,000+ Instagram followers** (top reel 653K views), 35K+ TikTok likes, and a 200+ video YouTube series, "A New York Diary."

SELECTED PROJECTS

Truth Layer, agent-honesty field study (2026)

Do autonomous agents faithfully report their own work? 69-day study instrumenting every session on a production automated machine: 7,186 claims across 4,330 sessions; of 5,025 witnessed against ground truth, 22.4% machine-confirmed and 1.2% contradicted. Finding: agent self-report is a poor oversight channel. Harness, writeup, and frozen aggregates published. github.com/MARCCHERGGI/truth-layer

NUMEN, production tool-using agent (2026)

Agent with persistent per-user memory and live news research over 4-round agentic tool-calling loops. Groq Llama-70B, ElevenLabs real-time voice, shipped as an installable PWA. Live, on a paid tier. numen-presence.vercel.app

JARVIS_HOME, desktop AI assistant (2026)

Cinematic Electron voice-briefing assistant with wake-on-clap, HUD, pluggable LLM brain, and multi-seat "Council" reasoning. 168 first-party TypeScript files, shipped as a 137 MB Windows installer, live-routes DeepSeek to Groq with failover. Electron 32, TypeScript, three.js. github.com/MARCCHERGGI/jarvis-home

World Cup Edge Finder, ML prediction and calibration (2026)

Dixon-Coles match-outcome model trained and backtested on 49,472 matches. Beats the naive baseline out of sample (RPS 0.207 vs 0.239, 90% bootstrap CI below zero) and stays calibrated (ECE 0.052).

Additional shipped work

Shift AI (shiftai-six.vercel.app), NYC job-search coach: chat agent, route planner, interview drills, iOS Capacitor shell, with rate limiting, circuit breaker, cache, and provider failover. **Scout** (scout-layer.vercel.app), a live product on the MCP and A2A interop standards. **Semantic Navigator**, client-side intent re-ranking with MiniLM in ONNX/WASM, no API key and no server inference. **Hergi Editor**, a local keyless video pipeline (Whisper ASR, LLM planner, ffmpeg) that cut a real 1080p YouTube episode end to end.

SKILLS

AI and ML: Production LLM agents, tool-use and function-calling loops, RAG and retrieval grounding, persistent agent memory, multi-provider routing and failover, prompt engineering and evaluation, real-time voice (STT, streaming LLM, TTS), on-device inference (ONNX, WASM, MiniLM), model calibration and backtesting, MCP and A2A tool authoring.

Languages and platforms: TypeScript, JavaScript, Python, SQL, Next.js, React, Node.js, Electron, three.js, Capacitor, Vercel, Postgres, Supabase, Stripe, Anthropic Claude, OpenAI, Groq, DeepSeek, ElevenLabs, Whisper.

Engineering practice: Rate limiting, circuit breakers, caching, CSP and secrets hardening, HMAC request signing, observability, CI and same-day deployment.

EDUCATION

High School Diploma

Self-taught applied AI engineer. Portfolio: marco-hergi.vercel.app