

Michal Malček

AI ENGINEER · SYSTEMS DESIGN · ADVANCED SEO

Poprad, Slovakia · [github](#) · [linkedin](#) · miso.malcek@gmail.com · +421 950 200 395

Summary

AI engineer building full-stack agentic systems end to end – Claude Code, Cursor, RAG, local and cloud AI. Delivered a Claude-based knowledge hub to a 15-person non-engineering team and handed it over clean: **1,352 tool calls at 93.9% success on ~\$30/month**, 50.2M input tokens, one 8 GB VPS. Then built a sovereign local-model system: **64,644 lines of TypeScript on 29 runtime dependencies**, because 12 things a library would have brought in were written by hand in 2,005 lines instead. **I measure before I claim** – every figure here has a reproducible path behind it. 20 years on both sides of software delivery (ERP, CRM, B2B, B2C, web, cloud), 8 of them leading SEO internationally across CEE, Gibraltar, Malta and Hamburg.

What I bring: architectures that survive handover · local AI where data must stay put · cloud and edge AI where scale earns it · retrieval that returns the right thing, measured · a feature built once and reused across systems rather than rebuilt.

Core strengths

- **Production LLM systems end to end** – Anthropic SDK, agent loops, tool calling, prompt architecture, handover to non-engineers
- **RAG that is tested, not assumed** – pgvector, Qdrant, Apache AGE, hand-written BM25+vector fusion (RRF), graph expansion, provenance tiering
- **Retrieval measured, not asserted** – behavioral rule retrieval went from **1 in 5 to 10 of 10** by storing rules as the situations that trigger them. The technique is weaker on tool selection (7/10) and I say so
- **Sovereign / local AI** – custom llama.cpp build, GGUF quantization, KV-cache tuning, prompt-cache economics; a 12B model doing agentic work unattended on one machine
- **Dependency discipline** – 12 modules hand-built rather than imported (JSON repair, HTML-to-text, RRF fusion, hybrid scoring, local mutex, context hot-swap, tool-call salvage, embeddings client, MCP bridge, event bus, session store) – fewer supply-chain surfaces and every failure traceable to code I own
- **Google Cloud APIs in production** – Search Console, Analytics Data (GA4), Knowledge Graph, PageSpeed Insights, CrUX, Cloud Storage, AI Studio, service-account OAuth
- **Edge computing** – Cloudflare Workers deployed by API, Workers KV, Workers AI, AI Search with query rewrite and reranking, authoritative DNS, SPF/DMARC
- **Cost engineering** – routing cheap models for tool selection and strong ones for synthesis; free-tier edge AI load-verified for production traffic. A 15-person deployment ran on one small VPS
- **Delivery across targets** – one codebase as a one-click Mac/Windows app, a single-VPS deploy, or behind a cloud endpoint
- **Frameworks evaluated against a real system, then not used** – LangChain, LangGraph, LlamaIndex, PyKEEN, FAISS. Each sells an abstraction I had already built and could instrument; PyTorch and TensorFlow are training frameworks and nothing here trains. **Where a gap is real I name it:** no cross-encoder reranker and no scored retrieval eval yet, both on the list

Public portfolio

Three shipped applications, source public, installers downloadable. Each is **architected to fit its target**, and the stack is chosen so the same architecture scales rather than being rewritten: near-zero dependencies, standard interfaces, no runtime that ties it to one vendor. The same codebase runs as a desktop app, on a single VPS, or behind any cloud – GCP, AWS, Cloudflare or a self-hosted cluster.

Crystal RAG · [downloads: macOS + Windows](#) · [repo](#) · [video to follow](#)

Point it at a folder, get a local searchable knowledge crystal. **Qdrant and the embedding model ship inside the app** – nothing to install, no data leaving the machine. Tauri + Rust, Mac and Windows.

Hybrid retrieval fuses vector, lexical and graph lists (RRF, k=60), so a file the vector search missed still surfaces through its relations. Optional endpoint for naming and chat, cloud or local, user's own.

The Wisdomkeeper · [downloads: macOS + Windows](#) · [video to follow](#)

AI-powered game with a **custom inference runtime written in Rust – 2,841 lines** over llama.cpp bindings (`llama-cpp-2`, optional Metal). Not a wrapper: a log-probability judge for decisions, a grammar-constrained actor so output is always valid, a supervised sidecar on stdio with nothing on the network, and two x86 build paths (AVX2+FMA and a `x86-64-v2` compat build) selected at launch. Models download on consent with pinned checksums. **Both models load and run on an 8 GB Windows laptop with integrated graphics** – tested. Desktop source private, installers public.

SEO Brain · [repo](#) · [release and video in progress](#)

Team Brain rebuilt as a product, **deployable three ways from one codebase**: single Linux VPS as it runs today, one-click macOS app, one-click Windows app. Express + Next.js + Anthropic SDK, 41 Postgres migrations. Credentials are the user's own, entered in settings: Claude API, OpenRouter or any OpenAI-compatible endpoint including a local one, plus Google service-account tokens for GSC, GA4, PageSpeed and CrUX. **The clearest demonstration that external connectivity works** – cloud model, local model and five Google APIs behind one settings screen.

Production systems – references

Team Brain · AI knowledge hub, msg life Slovakia · delivered and handed over Jun 2026 · [video to follow](#)

Claude API for a 15-person SEO and content team, on one Hetzner 8 GB VPS. Every figure below is from a database-generated usage report:

- **1,352 tool calls at 93.9% success**, across 25 distinct tools, each with its own measured rate – including the weak ones
- **~\$30/month total API spend** (billed; ran closer to \$22) – Haiku for tool routing, Sonnet for synthesis
- **15 users, all 15 active** · 446 conversations · 3,558 messages · 549 agent runs · 1,480 reports
- **50.2M input / 2.25M output / 4.95M cached tokens** over those 549 runs
- **1,121 articles · 514 memories · 1,454 embeddings** produced
- Four live data sources: Search Console, GA4, Ahrefs, CrUX. Seven-layer prompt system, agents, personas, semantic layer
- **Haiku for tool routing, Sonnet for synthesis** – a deliberate cost/quality split, in the code not the pitch
- Official Anthropic SDK from the first commit. Near-zero hallucinations, full documentation, clean handover to non-engineers

Factorium · sovereign AI system, running daily on one Mac · private, demo on request · [video to follow](#)

Gemma-4 12B Q5 doing real agentic work with no cloud inference. 69 native tools, unified retrieval over Postgres, pgvector, Qdrant, Apache AGE and TimescaleDB.

- **64,644 lines of TypeScript on 29 runtime dependencies** – "cut the middlemen" as a measurement, not a philosophy
- **Longest unattended run: 120 turns, 120 tool calls, 67 minutes**, no turn cap hit
- **477 logged tool calls · 0 reasoning failures** · all 63 errors were environment or guard, 75% recovered immediately
- **78,413 embedding points across 18 collections** · code graph of 3,567 entities and 11,002 relations over 5 repositories
- **6,169 organic memory facts** with real timestamps, retrieval at ~150 ms p50
- Research corpus with provenance tiering: 5,629 chunks, 187 files, every chunk tagged primary / own / synthesis / OCR
- **34× latency regression found and fixed** – traced to one disabled prompt-cache flag; 24.2 s → 1.0 s per task
- **Memory index repaired end to end**: 77% of the corpus was being silently truncated before reaching the vector store; 1,050 → 15,232 chunks with recall held at 8/10 rank-1 through 14.5× growth

Features move between these systems rather than being rebuilt – the RRF fusion, the embeddings client, the tool-call salvage and the session store were each written once and reused.

Experience

Owner · AI Engineer · SEO · Web

My Marketing, s.r.o. · Jan 2012 – present · Slovakia · Remote

- Build and ship the systems above; publish the measurements behind them
- **Cloudflare edge platform, two production zones** (aisvet.sk, alteranai.com) – a **Workers** edge proxy deployed via the REST API that instruments server-side pageviews with ``waitUntil()`` fire-and-forget, so analytics add zero latency to the visitor; Workers KV, Workers AI, authoritative DNS with GitHub/GitLab Pages subdomains, SPF + DMARC
- **aisvet.sk** – AI magazine on an own CMS with a **Cloudflare AI Search chat on the free tier**, load-verified for production traffic: hybrid retrieval, a query-rewrite model before search and a reranker after it, similarity caching. Same pattern going to a second site
- Custom **CDP browser automation** for Gemini and Google AI Mode where no API exists
- **Advanced hook systems** for Claude Code and Cursor, plus a custom operating mode for Claude Code – the apparatus that makes long autonomous sessions reliable
- Measurement write-ups on applied AI: rules retrieval and alignment, what actually limits a local model, RAG efficiency, entropy taxonomy – [full pack](#)
- 3D sales configurator with order CRM for custom car seats – [live since 2024](#)
- Prior: SEO rebrand of an online lottery outlet, lottery data bot, programmatic SEO, Google Partner agency (€10k/month ad spend)

Senior SEO Specialist → AI Engineer

msg life Slovakia · Sep 2024 – Jun 2026 · Remote

- **Traffic doubled in a year** on WordPress with slow release cycles – the constraint was the pipeline, not the content
- First large topic clusters and content pillars across technical (Java, Python, QA, SDLC) and HR domains
- Moved into AI Sep 2025, formalized as **AI Engineer** by Feb 2026; delivered and handed over Team Brain Jun 2026
- Moved the whole team from ChatGPT to Claude
- Numerous SEO and **4 AI workshops** for the entire Slovak branch, voluntary attendance, leadership to newest hires – the first opened with AI ethics and the practical limits of formal logic, before Team Brain existed as anything but an idea

Digital Media Manager

Swiss Re · Nov 2024 – Feb 2025 · Bratislava · Hybrid

- Sigma annual report readership **1,500 → 4,500**, plus other leadership topics (wildfires, climate risk) and product pages on comparable topics – a DR 90+ domain
- Cleared Jira tickets untouched for two years: CMS, DAM and Search Console settings fixed in a few clicks and one call
- 3 Ahrefs workshops; every team member reported new skills at quarterly review
- Two full-time roles held in parallel for three months, hybrid and remote, kept apart. **After probation I chose to continue only with the more suitable remote role** – two full-time jobs were not sustainable long term
- Completed all 10 mandatory trainings including AI Security & Compliance and AI Copyright

- First hands-on LLM work: ChatGPT Pro, loading articles into memory and feeding Ahrefs data in by hand – the problem Team Brain later solved properly

Search Engine Optimization Manager · DACH

LottoStarlet Ltd. → Tippland GmbH · Sep 2017 – Mar 2023 · Malta → Hamburg

- SEO strategy for a lottery startup across DE, AT, CH, UK: content, keywords, domains, link-building, Angular prerendering, agencies and freelancers
- Built embeddable, trackable affiliate ad units usable inside any article
- [Tippland Magazin](#) – built it, content included; **still ranking 9 years later with zero maintenance**
- Malta → Hamburg (2018) → remote from Slovakia (2019) → DE-only rebrand to Tippland.de (2022)

Search Engine Optimization Team Lead · CEE

Lottoland Ltd. · Jul 2016 – Aug 2017 · Gibraltar

- Managed 4 internal and 5 external colleagues across SK, CZ, PL; €15k/month link-building budget
- Created "Keyword Heroes", the internal Ahrefs training series; completed the 6-week StepUp management program

Online Marketing Manager

Individual Auto Design GmbH · May 2015 – Jun 2016

- Full-stack online marketing and in-house web development in 5 languages (DE, SK, CZ, HU, EN); WordPress + Magento, internal analysis for SAP Business One
- Launched every online channel from near-zero, migrated a content-heavy legacy site, then **grew visits and orders every month**

Google – two internship programmes

Online Media Associate Programme, EU HQ Dublin · 2011 · Associate Product Marketing Manager, Prague · 2010–2011

- Turned manual categorisation into an auto-improving customer database with scripting – presented to a Mountain View product manager, who **put a team on it**; delivered in weeks instead of two months
- 2× Peer Bonus for AdWords help guides; Google Student Ambassador; 3 co-marketing deals, verticalized Direct Mail (top 3 globally); on the team when Google passed 51% market share against Seznam

Sales Manager (from Junior Consultant)

Asseco Solutions · Sep 2006 – Dec 2009 · Prague

- ERP/CRM (Helios Orange) for SMBs: **14 successful implementations**, negotiated pilot deals in DE and PL
- Started as Junior Consultant and became the **youngest ERP/CRM sales manager in the company**

My AI stack

Models & serving Anthropic SDK · Claude API (Sonnet, Haiku) · llama.cpp (custom build) · GGUF quantisation · Gemma-4 12B · bge-m3 embeddings · HuggingFace · OpenRouter · Cloudflare Workers AI · Google AI Studio · OpenAI-compatible endpoints

Retrieval pgvector · Qdrant · Apache AGE · RRF hybrid fusion (hand-written) · knowledge graphs · semantic chunking · provenance tiering

Google Cloud Search Console · Analytics Data (GA4) · Knowledge Graph · PageSpeed Insights · CrUX · Cloud Storage · service-account OAuth

Languages TypeScript · Python · Rust · SQL · PHP

Web & desktop [React](#) · [Astro](#) · [Next.js](#) · [Express](#) · [Three.js](#) · [Tauri](#)

Data [PostgreSQL](#) · [TimescaleDB](#) · [Ahrefs API](#)

Edge & infra [Cloudflare Workers](#) · [Workers KV](#) · [Workers AI](#) · [AI Search](#) · [AI Gateway](#) · [DNS/SPF/DMARC](#) · [Nginx](#) · [PM2](#) · [VPS \(full setup\)](#) · [GitHub Actions](#) · [launchd](#) · [Docker-free by design](#)

Tooling [Claude Code \(custom mode, advanced hooks\)](#) · [Cursor \(hooks\)](#) · [Copilot](#) · [Git with textbook commit discipline](#)

How I work with AI

Measure before claiming – a number without a reproducible path does not ship. Full transparency is the best security. Textbook commits and documentation. Local-first: anything that can run on my machine does, so data does not travel when it does not have to. No plugins, no MCP, no skills files – everything explicit, auditable, behind one local endpoint. The operating line is *non ducor, duco pariter*: not being led, and moving together – sovereignty first, because without it the second is obedience with a nicer name.

Education

CEMS International Management (Master's, English) · 2009–2011 · avg 1.7 · exchange semester at Corvinus Budapest (A in all subjects) · best business project team (GE Money Bank) · full curriculum completed, not graduated · +41 ECTS in Information Systems & Technologies, avg 1.5

Applied Informatics (Bachelor's) · University of Economics Prague · 2006–2009 · avg 1.6 · top 10 of 500 in first year

Electrical Engineering, ICT · High School of J. Murgaš, Banská Bystrica · 2002–2006 · #1 in final year, avg 1.0

Startup Scholarship, Lisbon · 2013 · NOVA University + Startup Lisboa

Languages

Slovak native · **English advanced** (full master's taught in it, daily working use) · Russian A2 (3rd worldwide, St Petersburg University olympiad) · German A1 (Goethe-Institut Hamburg) · Spanish basics · Hungarian basics

Credentials

[All credentials, references, certificates and diplomas](#)

Open source & public output

[claude-skills](#) – the working apparatus behind a year of human-AI collaboration · [codex-symbiosis](#) · [atauri](#) · [gkgg](#) – Google Knowledge Graph API wrapper · [ai-alignment](#) · [aisvet.sk](#) AI magazine + [YouTube](#) + [podcasts](#) · [9 research notebooks](#) · browser games built one-shot with AI