Philipp Werdecker – KI, LLM-Agents und Voice

IT-Consultant und Entwickler, seit 2021 selbstständig. Schwerpunkt: LLM-basierte Backends, Agenten-Systeme und Echtzeit-Voice, gebaut auf Python-Plattformen, die bei Kunden seit Jahren produktiv laufen. Arbeitet in klar abgegrenzten Arbeitspaketen mit definiertem Ergebnis.

Rahmen: Remote (Fulda, Hessen), Vor-Ort-Termine nach Absprache · Deutsch, Englisch

---

Schwerpunkte

- LLM-Backends und Agenten: Orchestrierung, Tool-Anbindung (MCP), strukturierte Ausgaben
- Retrieval und Wissensabruf: pgvector, Qdrant, GraphRAG
- Evaluation auf echten Daten: Abnahmetests, Testdatensätze, Quellenbelege
- Echtzeit-Voice: Streaming, ASR, TTS, Latenzoptimierung
- Typisiertes, asynchrones Python (FastAPI, Pydantic) mit automatisierten Tests
- KI-gestützte Entwicklung mit Claude Code und OpenAI Codex im täglichen Einsatz

---

KI-Projekte (eigene Produktentwicklung, 2026)

RaySpec – KI-Backend-Plattform (Open Core, GitHub, npm, PyPI)

- Eine YAML-Datei erzeugt Auth, Mandantentrennung, HTTP-API, Postgres-Datenschicht, durable Jobs und Agenten
- Agenten über vier austauschbare SDKs: OpenAI Agents, Claude Agent SDK, Codex, Pi
- TypeScript, Node, Drizzle, Hono

Meeting-Intelligence (Memovo)

- Desktop-Aufzeichnung ohne Bot (Tauri/Rust, FastAPI, PostgreSQL), ca. 890 automatisierte Tests
- Auswertung mit Quellenbelegen auf Basis von LangGraph und pgvector
- Abnahmetests auf echten deutschen Transkripten bestanden

Voice-Agents

- Echtzeit-Sprachdialog (Deepgram, OpenAI, ElevenLabs) mit ca. 165 ms Median bis zum ersten Ton
- Eigener ASR-Server (Parakeet, MLX/CUDA)
- Eigenes deutsches TTS-Modell, trainiert auf einer Consumer-GPU

Runsera – Aufgaben- und Kontextverwaltung für KI-Agenten

- MCP-Schnittstelle, live auf runsera.com

Lokale LLM-Memory

- Dokumentwissen ohne RAG in ein lokales 14B-Modell auf 8 GB VRAM gebracht, Recall im Test von 33 % auf 100 %

Produktionserfahrung aus Kundenprojekten

IoT-Cloud-Plattformen (Hersteller Smart-Home-Geräte), seit 08/2021

- Zwei komplette Cloud-Plattformen für WiFi- und Zigbee-Geräte, sechsstellige Anzahl Geräte im Feld
- REST-APIs mit FastAPI und Flask inklusive Auth und User-Accounts, MQTT, PostgreSQL, MySQL
- Schnittstellen zu großen Handels- und Telko-Partnern, Alexa-Smart-Home-Skill (AWS Lambda)

Gebäude-Automationssystem auf LoRaWAN-Basis (Main Maintainer), seit 01/2026

- Backend auf ChirpStack-Basis, FastAPI async, Ports & Adapters, Firmware-Updates over the air

Enterprise-Infrastruktur (Engineering-Dienstleister Automotive), seit 2021

- Betrieb einer Citrix-Plattform für ca. 8.000 Mitarbeitende inkl. GPU-VDI auf Nutanix

---

Technologien

Python, TypeScript, FastAPI, Pydantic, Pydantic AI, LangGraph, Instructor, MCP, OpenAI- und Anthropic-SDKs, Claude Agent SDK, pgvector, Qdrant, Neo4j, FalkorDB, PostgreSQL, Deepgram, ElevenLabs, Parakeet/Whisper, MLX, CUDA, Tauri/Rust, Claude Code, OpenAI Codex, Git

---

Ausbildung

- 2018–2021: Fachinformatiker für Systemintegration (IHK), mit Auszeichnung

---