

Juan Cruz Fortunatti

Agentic AI Engineer · Agent Evaluation, LLM Infrastructure & Multi-Agent Systems

Berlin, Germany · +49 1529 2616523 · juancfortunatti@gmail.com
linkedin.com/in/juancfortunatti · github.com/apresmoi · ledeluge.me

I build agents and the harnesses that tell you whether they work. At Delivery Hero I built MATEvals, a production eval framework that replays cases from 10,000+ real agent traces with tool-call assertions, LLM-as-judge checks, and regression gates before rollout. Nine years shipping production software, three on LLM systems; previously led an R&D/innovation team and co-founded a voice AI startup.

EXPERIENCE

Delivery Hero | Berlin, Germany

Agentic AI Engineer · Dec 2025 – present

- **Production eval harness (MATEvals, Python):** replays and evaluates cases drawn from 10,000+ production agent traces; deterministic assertions on tool calls and arguments, LLM-as-judge on outputs, behavior-path comparison across versions, regression gates before rollout. Wired into CI: every PR my team ships is gated by a multi-sample MATEvals run (5 samples per case) over a curated set of CI/CD test cases.
- **Eval at scale on a laptop budget:** cut runtime memory to ~200 MB per worker, enabling 20–30 concurrent workers; feature suites run 20–40 cases × 5 samples (100–200 executions) to measure behavioral variance, not just pass/fail.
- **Multi-agent workflows in production:** build and operate vendor workflows with Pydantic AI and LangGraph (routing, tool use, context construction, structured outputs); shipped an agentic catalog-editing system for 50,000+ vendors that rewrites item names, descriptions, prices, availability and translations, with per-market feature toggles configured independently by each brand and market; ~2,000 vendor-accepted changes per month.
- **Observability → data flywheel:** spearheaded, designed and shipped a conversation visualization and annotation system for bot conversations across seven brands, now used by 275 internal users across Delivery Hero brands and markets, with role-based access and tenant isolation; labeled production conversations feed back into eval sets.
- **Reliability:** diagnosed production failures from oversized trace payloads and slow Langfuse writes; reduced trace payload volume by 97% while preserving eval-critical context.

Improving | US, remote

AI Solutions Architect & Lead Consultant · May 2024 – Dec 2025

- Architected a Python/LangGraph platform for building, testing and demoing MVP agents; cut project turnaround by 30% and initiated three client projects.
- Led technical discovery with clients: turned ambiguous requirements and data constraints into scoped agent workflows and delivery plans.

Leniolabs | US, remote

Head of Innovation · Jan 2022 – May 2024 · Tech Lead, PayPal engagement · Feb 2019 – Jan 2022

- Led the research team: GenAI prototypes and MVPs, an internal AI Guild translating new LLM techniques into engineering guidance; ran the Lenio Talks cycle (50+ talks) and conducted 70+ technical interviews.
- ML engineer for TriNet: internal assistant for call-center agents, end-to-end on AWS (data processing, prompt iteration, deployment, operations).
- PayPal: React data-visualization library adopted across 20+ analytics projects; Python pipelines and vector-tile generators for large geospatial data.

Also: Vocale.ai, Co-Founder & Lead AI Engineer (2023–2025, part time) – voice workflow platform piloted with Naturgy. Walt Disney Studios (2022) – data-viz dashboards for movie catalogs. Paradigma del Sur, Full Stack (2017–2019).

SELECTED PROJECTS

- **Coding-agent eval harness (PolyCSS / GlyphCSS):** runs coding agents in fresh workspaces, builds their output and grades the rendered result in Chromium, with a no-skill control and a Three.js baseline. Caught a grader that passed an invisible backface-reversed mesh; replaced DOM bounding-box checks with painted-pixel measurement and added mutation tests that verify graders catch injected bugs. github.com/layoutit/polycss/tree/main/eval/skill
- **Spawnfile:** spec and compiler for autonomous agents – compiles agent/team definitions to runtime configs and Docker images with explicit capability-support reports. spawnfile.com
- **Moltnet (Go):** self-hostable chat network for AI agents – rooms, DMs, persistent history, bridges for five agent runtimes. moltnet.dev
- **Jiang Lens:** agent-run research site that ingests a lecture corpus, grounds every claim to exact passages and timestamps, separates source from interpretation, and tracks forecast-style claims against outcomes. jianglens.com
- **Craft:** GlyphCSS, PolyCSS, css.earth, asciiQuake – 3D/ASCII rendering engines for the DOM, largely built with coding agents. PolyCSS ~770 and GlyphCSS ~210 GitHub stars.

TALKS

- **Harness the Harness** – Agentic Engineering Berlin, June 2026: building sensors that let coding agents see what they build. Oracle pattern: diff a ground-truth render against the output and hand the agent structured JSON of what's wrong; evaluate conversational agents across every node and tool call, not just input/output.
- 50+ internal talks organized and hosted as lead of Lenio Talks; external talks on React, D3, AI and data (incl. Nerdearla) – see ledeluge.me/talks.

SKILLS

Evaluation: trace replay, trajectory analysis, tool-call assertions, LLM-as-judge, regression gates, grader validation, mutation testing, annotation workflows, observability (Langfuse).

Agents: Pydantic AI, LangGraph, MCP, tool interfaces, structured outputs, routing, context construction, multi-agent orchestration.

Engineering: Python, TypeScript, Go, SQL, React, Playwright, PostgreSQL, Docker, Kubernetes, GCP, AWS, CI/CD.

Education: Physics coursework, Universidad Nacional del Sur, Argentina (2010–2015). **Languages:** Spanish (native), English (fluent).