

45.5×

monthly coding-agent
tokens vs the field median
(~3.96B / mo)

2.76×

orchestration success with
AgentEdge vs ReAct / LATs

425 /s

LLM inferences served on
NVIDIA Triton

99.99%

uptime on the production
MLOps platform

8 / 8

peer-reviewed papers, all
published

I design and evaluate LLM agent systems, then ship the path that runs them: inference serving, MLOps, and experiments that hold success fixed while cutting cost. AgentOne ran live at *Mobile World Congress 2026*.

01 EXPERIENCE

Senior AI Research Engineer · Nearby Computing 04/26 – now · Barcelona

- Owns **AgentOne** architecture; transitioned it to the development team and directed delivery.
- Mentored an intern shipping AgentOne’s frontend and MCP tooling from design to production.

AI Research Engineer · Nearby Computing 03/23 – 03/26 · Barcelona

- Built **AgentOne**, a multi-MCP agentic orchestration system for autonomous infrastructure management; presented live at **MWC 2026** with **300+** runs per day.
- Architected the **NBI MCP server**: 16 tools abstracting Kubernetes orchestration APIs for LLM agents.
- Developed **AgentEdge** (IEEE TMLCN): role-specialized agents with pre-execution validation, **2.76×** success vs ReAct/LATS; built the orchestration evaluation benchmark and scenarios.
- Engineered the LLM inference pipeline on **NVIDIA Triton** at **425 req/s**; **99.99%** uptime with failure-mode detection via Prometheus and Grafana.
- Designed ActSimCrit planning with digital twins; agent memory on **Milvus** with context engineering.

AI Research Engineer · Columbia University / AI4Finance 01/22 – 02/23 · New York

- DRL framework for live HFT market environments; **46%** less backtest overfitting (AAAI 2023).
- Contributor to **FinRL** (14k GitHub stars); grew engagement **214%** while moderating contributions.
- Led a team deploying parallel and distributed ML pipelines with MLOps on GCP.

AI Signal Theory Research Engineer · TNO 06/19 – 06/20 · Delft

- Control system for an optical satellite terminal: **86%** less control error; laser links beyond 1 Tbps.

Robotics Research Engineer · Swiss Plasma Center, EPFL 11/18 – 04/19 · Lausanne

- Redesigned a robotic control system, **91%** less control error; full FuseNet scholarship.

02 SELECTED PUBLICATIONS

Distributed Agentic AI Framework for Autonomous Edge-Cloud Service Orchestration
IEEE TMLCN 2026 · AgentEdge · doi:10.1109/TMLCN.2026.3705714

Attention-Driven AI Model Generalization for Workload Forecasting in the Compute Continuum
IEEE TMLCN 2025 · OmniFORE · doi:10.1109/TMLCN.2025.3584009

AERO: Adaptive Edge-Cloud Orchestration With a Sub-1K-Parameter Forecasting Model
IEEE TMLCN 2025 · doi:10.1109/TMLCN.2025.3553100

Deep Reinforcement Learning for Cryptocurrency Trading: Practical Approach to Address Backtest Overfitting
AAAI 2023 · arXiv:2209.05559

8 peer-reviewed, all published. Full list: berendgort.dev/research

03 EDUCATION

PhD, Artificial Intelligence

CUM LAUDE
Universitat Politècnica de Catalunya
10/23 – 03/26

8 publications. MareNostrum 5 HPC allocation for transformer training.

MSc AI & Computer Science

MAGNA CUM LAUDE
Harbour.Space University
09/20 – 12/21 · 9.3 / 10
Full scholarship. Thesis: PPO for multi-crypto trading (10 / 10).

MSc Mech. Eng., Control Technology
TU Eindhoven
09/18 – 06/20 · 7.3 / 10

04 TOOLKIT

AGENTS
LangGraph, MCP, Pydantic, OpenAI API, Cerebras, Milvus

ML
Python, PyTorch, deep RL, forecasting, evaluation design

INFERENCE & MLOPS
NVIDIA Triton, Kubernetes, Docker, MLflow, Prometheus, Grafana, GCP, Linux

05 ALSO

Co-founder & CTO, SUMMA

Real-time multi-agent product, built full stack with one co-founder; **10k** weekly active users.

Automated trading

Since 2016; built and ran bots across exchanges.

06 LANGUAGES

English, Dutch native
Spanish C1 · German B2

Off the keyboard: skiing, kitesurfing, motorcycles.