Diego ROJAS

Software Engineering and AI Research

rojasdiegopro@gmail.com • LinkedIn • +86 150 1079 2682 • Beijing, China
Speaks: French, English, ≈ Spanish, ≈ Chinese.

PROFESSIONAL EXPERIENCE

Deep Learning Research Intern September 2023 - Current
Zhipu AI, Large Models Research *Beijing, China*

- Developing a LLM-based Code Interpreter solution for our AI-assistants offering. Owner of the project from research to training to implementation to deployment.

Software Engineer Intern April 2022 - July 2022
Koyeb, Serveless Computing Provider *Paris, France*

- Overhauled the telemetry pipeline of our serverless product, for high request throughput and tenfold metrics ingestion capacity.
- Participated in developing high-impact customer-facing features like user application usage metrics.
- Performed platform maintenance work, upgrading tooling, data stores and services incurring no downtime nor impact on functionality.
- Identified and resolved performance bottlenecks in a complex distributed environment.

DevOps Engineer Intern July 2020 - December 2020
Instadeep, AI Decision-Making *Paris, France*

- Streamlined the research process by prototyping an internal cloud platform for managing the training of resource intensive deep learning models.
- Conducted surveys with the research team to identify and address painpoints in the research and development process.

EDUCATION

Tsinghua University September 2022 - Present
Computer Science, Advanced Computing Master, GPA: 3.94 *Beijing, China*

- Relevant coursework: Big Data Systems (A), Deep Learning (A), Distributed Systems (A-), Natural Language Processing (A-).
- Thesis: Reinforcing Large Language Models of Code

Epitech, European Institute of Technology September 2019 - July 2023
Computer Science, Information Technology Expert, GPA: 3.80 *Paris, France*

- Relevant coursework: Systems Programming (A), Networking (A), DevOps (A).

AREAS OF EXPERTISE

Topics: Cloud Computing · Large Pre-Trained Models · Systems & Network Programming · DevOps/AIOps · Distributed Systems · Accelerated Computing · Front-End Development · Observability
Technologies: Go · C/C++ · Python · Kubernetes · Docker · Linux · Javascript · Rust

PROJECTS

Gopilot: Pre-trained a 290M parameters GPT-style model using a custom-built tokenizer for Go code. Relative +37% improvement on HumanEval. Open-sourced recipe, weights and extension.
42sh: Re-implementation of a Linux shell with support for 20+ shell features including complex scripting and job control in pure C.