

Pablo Arango

📍 Medellín, Colombia (UTC-5) | ✉️ juparangopl@gmail.com | 🌐 pabloarango.dev
🌐 linkedin.com/in/monsieurpablo | 📄 github.com/monsieurpablo

Senior AI Engineer · Python Backend · ML & Simulation Platforms

Summary

AI engineer and Python backend engineer who takes ML models and cloud services from prototype to production; 7 years of professional Python. At Hoare Lea, shipped ML inference services 20-100x faster than the physics simulation (internal users grew from 5 to 50+ engineers); now architecting an Azure platform for long-running simulations (FastAPI, IaC, CI/CD, OpenTelemetry, React + TypeScript). Builds LLM and vision pipelines with evals on a self-hosted platform; works agent-native daily (Claude Code, Codex, MCP).

Skills

Programming: Python (7 years), TypeScript, SQL, Bash

AI & Agents: LLM APIs, AI agents (tool calling, persistent memory), LLM evaluation (blind-label validation, per-run cost tracking), multi-agent workflows, MCP, prompt and context engineering, Claude Code, Codex, n8n, workflow automation

Backend & Cloud: FastAPI, REST APIs, PostgreSQL/PostGIS, Redis, Docker, Microsoft Azure (Container Apps and Jobs, Storage Queues and Blob, Azure ML, GPU VMs, Entra ID), Bicep infrastructure as code, CI/CD, OpenTelemetry, pytest, Playwright, React, Vue.js, Streamlit, Linux, Git, AWS and Terraform (coursework)

Machine Learning: PyTorch, LightGBM, graph neural networks, CNNs (U-Net), feed-forward networks, model serving (Azure ML, FastAPI), batch training-data pipelines (20k-100k cases per model), Pandas, Plotly

Simulation: Ladybug Tools, Radiance, EnergyPlus, EDSL TAS, OpenFOAM, TRNSYS, Rhino/Grasshopper, Rhino Compute, Speckle, geospatial (GeoPandas, Rasterio, LiDAR), CIBSE TM59 / Part O

Languages: English (full professional), Spanish (native), German (elementary)

Experience

Senior AI Engineer · ML & Simulation Platforms

Jun 2025 – Present

HOARE LEA (UK ENGINEERING CONSULTANCY) · CONTRACT

Remote, Medellín

- Architecting a cloud job platform on Azure for compute-heavy simulations: FastAPI job API, queue-triggered workers that autoscale to zero, object storage plus a SQL job store, React + TypeScript front end. Live in staging and production (daylight and energy first).
- Agent-native delivery: writes the specs, context and acceptance criteria; Claude Code and Codex agents implement; owns architecture, code review and verification, with automated gates on agent output.
- Productionized three ML inference services (FastAPI, Docker, Azure): internal users grew from 5 to 50+ engineers, now serving 500+ design iterations a month inside engineers' CAD tools.
- Reliability for jobs that run up to 2+ hours: schema validation rejects bad payloads in seconds, a visibility timeout plus heartbeat stops a long job being picked up twice, failed jobs return logs and a readable error, outputs are checked against a reference standard.
- Delivery: infrastructure as code (Bicep), CI/CD with staging on merge and gated production releases, Entra ID SSO and API keys, OpenTelemetry tracing across API, worker and UI, Playwright end-to-end tests.
- Built a GPU ray-tracing daylight service on Azure GPU VMs: typical-floor results in under 4 seconds, used on 8 live projects.

Senior Computational Developer & Machine Learning Engineer

Jan 2024 – Jun 2025

HOARE LEA · FULL-TIME

London, UK

- Built the batch simulation pipelines that generated training data: parametric sweeps over EDSL TAS, EnergyPlus and Radiance producing 20k-100k labeled cases per model, validated against the industry reference tool (IES VE) before scaling.
- Overheating model (graph neural network): led data generation, training and deployment. 97% accuracy, under 10 seconds per prediction versus hour-scale runs (~100x faster); in production use on 10+ client projects.
- Daylight model (U-Net CNN, architecture co-designed with a colleague): owned the 100k-case dataset and the Azure ML inference stack. ~3% MAE, ~20x faster than ray tracing; in production use on 30+ client projects.
- Peak solar load model (feed-forward network): 98% accuracy, ~1 minute per prediction, 50x faster.
- Serving and integration: Azure ML endpoints and FastAPI services in Docker, called from a Grasshopper plugin so engineers got results inside their CAD tool in seconds, with request batching to meet latency targets.

Application Developer

Jan 2023 – Jan 2024

HOARE LEA · FULL-TIME

London, UK

- Built and deployed an authenticated internal web app (Python, Streamlit, Speckle API, Heroku) used by 10+ consulting teams, cutting report production from 2 days to 2 hours.
- Automated data pipelines and dashboards (Pandas, Plotly) that standardized reporting across 25+ projects and cut manual data processing by ~75%.

Sustainability Consultant

HOARE LEA · FULL-TIME

Sep 2021 – Jan 2023

London, UK

- Built reusable parametric simulation pipelines (Python, Ladybug Tools SDK, Pollination cloud, Speckle) used on 20+ client projects.

Junior Project Engineer

TRANSSOLAR KLIMAENGINEERING · FULL-TIME

Mar 2020 – Sep 2021

Stuttgart, Germany

- Developed outdoor-comfort capability in TRNSlizard, Transsolar's in-house simulation plugin (GhPython in Rhino/Grasshopper), coupling a CFD solver (OpenFOAM) with a thermal solver (TRNSYS).
- Ran a multi-objective optimization loop (NSGA-II) on a 64-core server: 20x faster parametric thermal studies.
- Analyzed a 6M-row time-series database from ~1,650 building sensors (Python, Plotly) for building-automation commissioning.

Parametric Building Physics Consultant

SELF-EMPLOYED

Jan 2019 – Mar 2020

Medellín, Colombia

- Built a Python sensitivity-analysis framework that batch-ran daylight (Radiance) and thermal (EnergyPlus) simulations over 1,000+ design permutations, isolating the 4 parameters behind 85% of performance outcomes.

Personal Projects

Property Valuation Engine

PERSONAL PROJECT · LIGHTGBM, POSTGRESQL/POSTGIS, PYTHON

May 2026 – Present

Medellín metro area

- Fair asking price per m² for every home for sale in the Medellín metro area, with a calibrated interval and the comparable listings behind it.
- LightGBM quantile models, conformalized per property type: ~10-11% median absolute percentage error under spatial cross-validation; p10-p90 intervals cover 80% of held-out listings, as designed.
- Public listing data from 11 source adapters in one PostgreSQL/PostGIS schema (14 Alembic migrations); deduplication turns 50,456 active sale-apartment listings into 29,937 units; spatial joins add 32 zoning, hazard and transit layers.

LiDAR2Building API

PERSONAL PROJECT · FASTAPI JOB SERVICE AND CLI

Mar 2023 – Present

- Turns one latitude/longitude in England into extruded 3D context geometry (buildings with LiDAR heights, terrain, trees, streets); taken from a Python CLI to an authenticated job service other apps call.
- FastAPI job API (submit, poll, stream progress over Server-Sent Events, download, cancel) with idempotency keys and a result cache. arq worker on Redis, jobs and events in PostgreSQL, a heartbeat feeds the readiness check, orphaned jobs are failed and refunded.

Self-Hosted LLM Automation Platform

PERSONAL PLATFORM · LLM AND VISION PIPELINES WITH EVALS, AN AGENT AND COST CAPS

Feb 2025 – Present

- Evals: an LLM classification pipeline over 4,100 documents, validated on a 50-document blind sample labeled by a stronger model (92-100% agreement on key fields), \$0.73 total inference cost.
- The blind check flagged one biased field (74% agreement); the headline metric moved to fields with 100% agreement.
- Model choice by benchmark: a local 8B vs a hosted 32B vision model on the same 20 photos. Local: 28.5 s per photo, 5 schema failures. Hosted: 21.5 s, 1 failure, under \$0.0005 per photo.
- Shared long-term memory for my coding agents (Claude Code, Codex) on PostgreSQL + pgvector with local embeddings and reranker; migrated 18,268 records, secrets redacted before any API call.
- Self-hosted Hermes agent with tool calling and persistent memory.

Certifications

Certificates: Terraform, Platzi (course certificate) · AWS, Platzi (course certificate) · DGNB International Consultant (2020)

Education

Professional Degree in Architecture (5 years, 330 ECTS equivalent), Summa Cum Laude

UNIVERSIDAD NACIONAL DE COLOMBIA

2014 – 2019

Medellín, Colombia

- Graduated 1st in class, GPA 4.6/5.0. Thesis: Python sensitivity-analysis pipeline (SALib) driving EnergyPlus and Radiance simulations.

Academy Fellowship (TSA7), fully funded

TRANSSOLAR ACADEMY (TRANSSOLAR KLIMAENGINEERING)

Jan 2020 – Mar 2020

Stuttgart, Germany

- 400 hours of climate-engineering training: building physics, comfort, daylight and thermal simulation, CFD, energy-supply systems.

Awards & Publications

Awards: Best Paper, I+D+i Congress 2019 · Summa cum laude, 1st in class, UNAL 2019 · Academic Excellence Scholarship (8 times)

Publications: 7 peer-reviewed conference papers in 5 countries (2018-2020): PLEA 2020, ENCAC 2019 (3), I+D+i 2019, MACDES 2018 (2)