Walter J. Troiani Vargas

Computer & Data Scientist

Specialized in AI & Natural Language Processing

+34 642 35 60 26 | waltertv02@gmail.com | https://www.linkedin.com/in/walterjtv | https://ezwalt.github.io/

Education

Technical University of Catalonia (UPC) — Barcelona, Spain
B.Sc. in Computer Science — *Sep 2020 – Jun 2024*

Technical University of Catalonia (UPC) — Barcelona, Spain
M.Sc. in Data Science — *Sep 2024 – Jul 2025*

- **Nova 111 Student Spain 2026**, selective national top 0.01% CS talent award for academic and technical excellence.

University of Padova (Università degli Studi di Padova) — Padova, Italy
International Double M.Sc. in Data Science — *Sep 2025 – Jul 2026*

Experience

MLOps Engineer — February 2026 - Present
Multiverse Computing — *Barcelona, Spain*

- Spearheaded the development of an **AutoML framework for distributed LLM compression**, integrating distillation, pruning, and quantization into unified pipelines to enable no-code compression while maintaining within 5% of baseline performance across multiple benchmarks. Standardized and centralized company-wide compression recipes, **reducing experiment reproduction time from months to days.**
- Managed and scaled heterogeneous multi-cloud Kubernetes infrastructure supporting 1,000+ GPUs (H200, L40S, B200/B300), and **led deployment of one of Europe's first large-scale B300/B200 clusters (224+ GPUs)**, enabling distributed training and high-throughput experimentation through Coder and SkyPilot.
- Designed and **migrated 500+ TB** of ML storage to a modular multi-region architecture, provisioning isolated FSx and S3 resources via Terraform and CI/CD automation. Reduced storage footprint by 3× and infrastructure costs by 67% while enabling zero-downtime adoption across teams.
- Reconstructed end to end the **LittleLamb** training pipeline for sub-billion-parameter reasoning LLMs, covering pretraining, mid-training, and post-training stages with online off-policy knowledge distillation and supervised fine-tuning on large-scale distributed clusters. Achieved up to **50% model compression** within Megatron-LM while maintaining competitive reasoning performance for edge AI.

AI Research Assistant — February 2026 - Present
Telefonica Research — *Barcelona, Spain*

Researching advertising opportunities and optimal ad allocation policies in multi-turn LLM-based conversational agents focusing on optimizing the trade-offs between user trust, advertiser utility, and long-term platform sustainability. Building an experimental framework combining behavioral interaction data and psychometric features (OCEAN) for adaptive conversational recommendation, and proposing alternative evaluation metrics beyond CTR to capture trust-aware and long-horizon outcomes. Ongoing work includes the design of a multimodal user study using EEG and eye-tracking to measure neurophysiological responses to persuasion and different ad formats.

AI Engineer Intern — June 2025 - September 2025
Conseil Européen pour la Recherche Nucléaire (CERN) — *Geneva, Switzerland*

- Led end-to-end development of AccGPT RAG for BE-ICS support, boosting Answer Correctness by **+55%** with RAG integration and and iterative prompt engineering across 10+ foundation models, improving reliability for 250+ SCADA applications.
- Architected a **Scalable** MLOps ETL pipeline (CLI + Streamlit GUI) with 99.95% fault tolerance to ingest, summarize, embed and index ∼40M tokens of Jira cases (≈33× Shakespeare Corpus) into ChromaDB at 5M tokens/hr.
- Reduced Jira history size **10**× via a Map-Reduce hierarchical summarization algorithm for semantic retrieval at scale.
- Designed and implemented an automated, modular LLM-as-a-Judge evaluation framework with 7 custom metrics (Correctness, Faithfulness, Context Precision...), enabling cross-model benchmarking and cutting evaluation time by **orders of magnitude**.
- Automated Q&A test-set creation with human-in-the-loop curation, increasing expert throughput **1167%** and producing high-quality test data for benchmarking.
- Developed a Jira Issue Recommender leveraging semantic embeddings and HNSW algorithm to surface related historical tickets to cut duplicate troubleshooting.

LLVM Software Engineer — October 2024 - July 2025
Barcelona Supercomputing Center (BSC) — *Barcelona, Spain*

- Led the implementation of OpenMP 6.0 Loop Transformations in the LLVM community, including development of advanced optimizations such as Loop Fusion and Loop Fission in Clang's front-end to improve scalability in HPC applications.
- Actively contributed to the **OpenMP and LLVM communities**, managing complex git workflows, reviewing patches and mentoring new contributors in open-source development.

Data Scientist July 2023 - July 2024
MANGO *Barcelona, Spain*

- Developed an end-to-end LLM document translation application gathering requirements with the translation team on a weekly basis, **reducing costs by €30K in the first month**
- Created REST endpoints using Dockerized Lambda's, API Gateway, and S3 for a critical size curve prediction model, **cutting deployment time by 40%**
- Implemented a **recommender system** for automatic product descriptions using **Siamese networks** and the FashionCLIP model
- Crafter an advanced Streamlit dashboard with **explainable-AI (SHAP)** and KPI visualization to **enhance business decision-making in best seller predictions**
- Authored comprehensive knowledge-transfer and API documentation using OpenAPI, streamlining project onboarding
- Provisioned AWS services with Terraform through Jenkins CI/CD pipelines and BitBucket webhooks, adhering strictly to **DevOps and MLOps** methodologies and **led Databricks, Airflow and Python migrations on established, high-impact projects.**

Side-Career

External relations Manager October 2022 - June 2025
FibVisiona *Barcelona, Spain*

- Initiated outreach with over 50+ **leading tech companies** such as Mango, CaixaBank, HP, Dynatrace, Glovo...
- Authored a Battle Pass system to gamify the fair, enhancing user retention and engagement by atleast 33%
- Successfully **connected 2000+** students with tech industry opportunities and and managed external relations team of 3 members

Undergraduate Hardware Teacher October 2021 - June 2023
UPC *Barcelona, Spain*

- Achieved a 98%+ pass rate among students
- Taught over 100 students and conducted classes with over 40 students

Certifications & Licenses

Stanford Machine Learning Specialization	2023
DGT Driving License	2024
Professional Scrum Master I	2024
HuggingFace Model Context Protocol Course	2025
HuggingFace Agents Course	2025

Technical Skills

Languages:*C*, *C++*, *Python*, *Rust*, Java, SQL, JavaScript, HTML/CSS, R, Haskell, Prolog, Kotlin, Swift, Groovy, CUDA, Mojo, x86, RISC-V
MLOps: Git, Jenkins, Docker/Compose, Podman, Kubernetes, Helm, Airflow, AWS, Terraform, VS Code, Postman, FastAPI, HuggingFace, ZenML, CometML, MLFlow, Ragas, DeepEval, LlamaIndex, LangChain, LangGraph, Gradio, PydanticAI, CrewAI, DSPy, OLlama, vLLM, ArgoCD, SkyPilot, Coder, ClearML, Ray, Megatron, NeMo
Frameworks: Streamlit, FastAPI, Flask, Django, PySpark, PyTorch, Tensorflow, Scikit-learn, OpenCV
Compilers : LLVM IR, MLIR, Clang, Flang, OpenCL, OpenMP, OpenGL, GLSL, WebGPU, ONNX, TVM, GDB/LLDB
Databases & Big Data: Neo4J, PostgreSQL, OracleSQL, MySQL, DuckDB, Hadoop MapReduce, Spark, HBase, ZooKepeer, HDFS, Delta-lake, MongoDB, ElasticSearch, Talend ETL, Kafka, Amazon S3/RDS, GraphQL, GraphDB, SPARQL, Cassandra, ChromaDB, MinIO, Qdrant, FAISS
Project Management: Agile, Scrum, LeSS, Kanban, Extreme Programming (XP)

Languages

Spanish: Native

Catalan: Native

English: Advanced

Italian: Intermediate (B1)

German: Intermediate (B1)

Projects

Pedantically Reinforced Large Language Model (PRLLM): Pedantic SmolLM2 Preference Alignment with PPO based RLHF

Size does not matter: Sub-Billion VLM NanoChimera is All You Need!: Visual Pretraining of a Vision Language Model

Read My Mind: EEG Motor Imagery Decoding with the Multi-Scale Dual-Axis Conformer: Novel Dual Conformer

VibeRadar: End-to-End Live Product Sentiment-Analysis

eZAutoML: eZAutoML Repository PyPI Download

LLM from Scratch (WIP): LLM pretraining from Reddit

ClangIR - LLVM OpenSource contribs. (Thesis): Repository

The Troiani Programming Language (WIP): Repository

ZeOS custom Linux Kernel: ZeOS UNIX Kernel