Alberto Sánchez Pérez

Data Scientist - GenAI & NLP

MSc Data Science | BSc Mathematics and Computer Science

Profile

Data Scientist and AI Engineer specialized in NLP and GenAI. I bridge applied research (NAACL'25, SEPLN'23) with industry delivery, leading Fortune 500 projects and owning end-to-end data/AI pipelines (Ingestion, preprocessing, model tuning, deployment). Proficient in genai technologies, classic ML and MLOps. Effective cross-team communicator and fast learner in high-stakes environments.

Skills

GenAI: LLMs, Fine-Tuning (Hugging Face), RAG, Agents

AI: ML (supervised & unsupervised), DL (PyTorch, Transformers)

MLOps: Data Engineering (SQL, Alembic, DBT), CICD (Git, Docker), Cloud (AWS), Deployment (Airflow, FastApi)

Languages: Spanish (native), English (CAE - C1), French (DELF - B1)

Work Experience

Aily Labs — **Madrid, Spain**

Mid Data Scientist — *Aug 2025 – Present*

- Delivered GenAI features for Fortune 500 clients: Large-scale LLM document extraction for legal contracts and entity matching across data sources.

Junior Data Scientist — *Sep 2024 – Aug 2025*

- Designed procurement document-processing pipelines, replacing RAG approaches with a MapReduce design that improved reliability and scalability.
- Led development of a Competitive Intelligence app (P&L forecasting, market insights, asset-buy recommendations). Database and pipelines design and orchestration, reducing runtime from 60 minutes to under 1 minute.
- Mentored an intern, coordinated sprints and code reviews, and regularly presented progress to C-level stakeholders.

Research Assistant — *Mar 2024 – Sep 2024*

- Main author of NAACL 2025 paper on using agent-based systems to generate insights from relational databases. (arXiv:2503.11664) Deployed to production and reduced operational costs by 99.6% on production pipeline.

Ontology Engineering Group (UPM) — **Madrid, Spain**

Research Assistant — *Sep 2021 – Jul 2022*

- Co-author on synthetic data generation paper from KG data (WidAug), boosting F1 by 20% for low-resource NER.
- Developed a combined NER and machine translation system for low-latency applications, reducing costs by 40%.

Publications

NAACL 2025: Main author: An LLM-Based Approach for Insight Generation in Data Analysis (arXiv:2503.11664).

SEPLN 2023: Co-author: Data augmentation for named entity recognition using Wikidata (WidAug)

Education

Eurecom — **Sophia Antipolis, France**

MSc in Data Science — *Sep 2022 – Jul 2024*

Polytechnic University of Madrid — **Madrid, Spain**

BSc in Mathematics and Computer Science — *Sep 2018 – Jul 2022*

Personal Projects

StopSlop: Chrome extension for low-quality site detection: Automated web scraping, custom hyperefficient FastText re-implementation in NumPy for real-time classification.