

Hang Yao

Barcelona, Spain hangyao2002@gmail.com +34 609 51 86 47
linkedin.com/in/hangui11 github.com/hangui11 resume.hangui11.com

Professional Summary

Data & Software Engineer with a BSc in Computer Science and an MSc in Data Science, plus enterprise experience at NTT DATA building production microservices for a banking client. Comfortable across the full development lifecycle — REST APIs, relational databases, automated testing and CI/CD in an agile team — and across modern data work, from scalable pipelines and cloud warehouses to machine learning and AI systems. Strong on turning complex problems into reliable, deployed solutions, with a habit of validating and measuring what I build.

Education

Sep 2025 – Jan 2026

Beihang University (China's top-10 in engineering) - *Master's Exchange Student*
- **Coursework:** Data Mining, Foundation Models, Information Retrieval.

Sep 2024 – Jun 2026

Universitat Politècnica de Catalunya (UPC) - *Master in Data Science*
- **Thesis:** Vision-Language Models (VLM) for Graph Extraction.

Sep 2020 – Jun 2024

Universitat Politècnica de Catalunya (UPC) - *Bachelor's in Computer Science*
- **Thesis:** State of the Art Recommender Systems (Collaborative Filtering & DL).

Experience

Feb 2024 – Jun 2025

Back-End Engineer Intern - *NTT DATA*
- Developed enterprise-grade microservices and state-driven workflows with the Spring ecosystem to securely route and process financial records for a major banking client, working across design, development, delivery and maintenance.
- Designed and consumed REST APIs between services, modelling application state onto custom relational database schemas to guarantee reliable data persistence.
- Wrote and maintained unit tests with JUnit, sustaining ∼90% coverage on core modules, and set up CI/CD pipelines with GitHub Actions; diagnosed and resolved production incidents under strict financial-data compliance.
- Worked in an agile team with code review and version control (Git), and modernised legacy multi-layer frameworks without disrupting existing functionality.

Projects

GitHub Link

Big Data Management (End-to-End Data Lakehouse)
- Engineered a containerized 3-tier data architecture (Cold/Warm/Hot) in Docker, integrating 10GB of diverse datasets via REST APIs using Spark and Redis, with Delta Lake as the storage layer.
- Orchestrated complex daily ETL workflows with Apache Airflow and built a web dashboard consolidating cross-platform metrics on top of the structured data lake.

GitHub Link

Event Analytics Platform (Streaming Ingestion to Cloud Warehouse)
- Designed and deployed two services end to end on Cloud Run: a Node.js/Express ingestion API that validates, enriches and publishes incoming events to Pub/Sub, and a Python/FastAPI analytics API serving metrics over HTTP from BigQuery.
- Built in reliability and cost controls: rejected malformed records at the edge, deduplicated at-least-once deliveries at query time, and added caching plus pre-aggregation that cut a representative query from ∼3 MB scanned to ∼2.4 KB.

GitHub Link

Steam Games Satisfaction Prediction & MLOps Pipeline
- Carried out data profiling and PCA feature selection on >150k API records, designing a custom Neural Network to capture non-linear relationships and maximize predictive R^2.
- Built an MLOps pipeline applying DuckDB across transformation stages for complete data lineage and instant state rollback, with a custom UI for automated execution, process monitoring and result visualisation.

GitHub Link

Review Summarizer with Retrieval-Augmented Generation (RAG)
- Built a RAG pipeline over large-scale review data: sentence-transformers embeddings + FAISS dense retrieval feeding a 4-bit quantized Qwen3-4B-Instruct generator for efficient local inference.

- Evaluated the system rather than assuming it worked: benchmarked against TextRank and BART baselines using ROUGE and an LLM-as-a-Judge protocol scoring accuracy, faithfulness and relevance; RAG scored 9/10 where baselines hallucinated.

GitHub Link

Floor Plan to Housing Graph Extraction Using VLMs (MSc Thesis)

- Built an annotation-free pipeline converting 1,276 raster floor plans into structured topological graphs using a frozen Vision-Language Model (Qwen3-VL-8B), with no fine-tuning or labelled data.
- Designed a deterministic validation framework that automatically checks outputs for structural consistency and flags anomalies, plus a persistent-memory refinement loop; 90% classification accuracy and structurally valid outputs for 90.2% of plans.

GitHub Link

STGF-LLM: Time-Series Forecasting & Synthetic Data Generation

- Fine-tuned a Large Language Model (Qwen3) for time-series forecasting and high-fidelity synthetic data generation, with a history-aware prompting strategy to serialize multivariate temporal data.
- Benchmarked against state-of-the-art generative models (TimeGAN, Chronos-T5), achieving competitive Train-on-Synthetic-Test-on-Real (TSTR) accuracy within hardware constraints.

GitHub Link

Multivariate Analysis of 167 Champions (R)

- Analyzed 167 game characters across base and growth statistics in R, applying PCA, MDS, MCA, cluster and discriminant analysis to recommend the best fit for each role.
- Delivered a reproducible R Markdown report with ggplot2/FactoMineR visualizations translating statistical findings into clear recommendations.

Technical Skills & Languages

Data Engineering: Data Pipelines (batch & streaming), Apache Spark, Apache Airflow, ETL/ELT, Delta Lake, Data Lakes & Warehousing, Pub/Sub, Data Quality, Governance & Lineage, MLOps.
Cloud & Databases: GCP (BigQuery, Pub/Sub, Cloud Run), Docker, SQL, Relational Databases, DuckDB, Redis.
Programming: Python, SQL, Java, R, JavaScript, HTML/CSS.
AI & Machine Learning: PyTorch, TensorFlow, Scikit-Learn, Hugging Face, Pandas, NumPy, FAISS, LLMs/VLMs, RAG, model evaluation (ROUGE, LLM-as-a-Judge), PCA.
Backend & Web: Spring Boot, FastAPI, Node.js, Express, Vue.js, REST APIs, Microservices.
Tools & Practices: Git, GitHub Actions, CI/CD, JUnit, Unit Testing, Agile.
Analytics & Visualisation: Tableau, Data Visualization, Time-Series Forecasting, FactoMineR, ggplot2, tidyverse.
Languages: Chinese (Native), Spanish (Fluent), Catalan (Fluent), English (Fluent).