

Ari Nubar Boyacıoğlu

Data Engineer | AI Tooling | Natural Language Processing for Low-Resource Languages

Karlsruhe, Germany · me@arinubar.com · arinubar.com · github.com/AriNubar · huggingface.co/AriNubar · linkedin.com/in/ari-nubar-boyacioğlu

SUMMARY

Data engineer with an M.Sc. in computer science from KIT, specialised in artificial intelligence and natural language processing. Building production data pipelines and financial data models since 2025; alongside that, author of the first neural machine translation system and parallel corpus for Western Armenian, an endangered language with no state support. Comfortable owning a problem from raw source system to the number a customer reads, and equally comfortable where the dataset has to be built before anything can be trained. Works with agentic development tooling and self-hosted models daily, and builds that tooling as well as using it.

Core: Python · SQL · Polars · Azure · Dremio/Iceberg · data pipeline architecture · financial data modelling · PyTorch and Hugging Face · LLM and agent engineering · German (C2)

EXPERIENCE

Data Engineer — ValueWorks GmbH

Karlsruhe, Germany · January 2025 – present

- Own end-to-end data pipelines — collection, transformation, KPI calculation — on a multi-tenant financial planning and reporting platform used by CFOs, boards and investors. 55+ integrated source systems spanning accounting and ERP, CRM, billing and subscription management, HR and banking, including Salesforce, HubSpot, Personio, Stripe, Zuora and Microsoft Business Central. The platform serves 90–100 active customers; I have shipped directly for around 30 of them.
- 146 Jira tickets delivered end to end, from requirement through implementation to production, and 490 merged pull requests between February 2025 and July 2026, with work shipped in every one of those eighteen months.
- Model financial data for reporting and planning: group consolidation, intercompany elimination, adjustment levels, deferred revenue recognition, budget against actuals and cash-flow statements, translated from group consolidation systems and accounting ledgers into customer-facing dashboards and drill-downs.
- Built the liquidity-planning and expense-management engine over raw accounting exports, the longest single workstream of my time there: clustering noisy free-text booking descriptions into normalised vendors, deriving contracts from the resulting groups, classifying payment cadence from transaction sequences, and projecting each contract's next expected payment to determine whether it was still running or had churned. Replaced manual revenue and expense classification and drives a rolling cash forecast.
- Designed the per-customer data-products layer: curated analytics-ready tables on a Dremio/Iceberg lakehouse plus a read-only API with manifest, endpoint-scoped tokens and snapshot time travel, giving customers governed access to their own transformed data, with client quickstart documentation.
- Reduced peak memory of the platform's largest pipeline (400+ KPIs, ~30 GB staging data) from over 50 GB to approximately 13 GB using streaming I/O, monthly partitioning and lazy evaluation in Polars, keeping it within its Azure Container Instance budget.
- Moved from taking requirements through Customer Success to running specification and troubleshooting calls directly with customers' finance teams, largely in German, turning their reporting logic into pipeline specifications and auditing KPI discrepancies across four pipeline layers against their live systems.
- First on the team to adopt AI coding assistants and, later, agentic development tooling; built the internal tooling that made them work against our codebase and drove adoption across engineering. Agent onboarding documentation with auto-generated code and KPI indexes, a static KPI dependency graph powering an "affected KPIs" pull-request review hook, reusable agent skills for tracing and documenting KPIs, and autonomous Jira-ticket-to-pull-request workflows, now part of how the team works. Personal delivery roughly doubled over the same period, from around 20 merged pull requests a month to around 40.

IT and Language Consultant — Calouste Gulbenkian Foundation

Armenian Communities Department, Lisbon (remote) · March 2025 – present

- Evaluated grant proposals in language technology, education and digital culture across the department's funding cycles, producing written assessments with funding recommendations for the director.
- Drafted the department's multi-year strategy for Western Armenian in digital environments, including comparative case studies of Basque, Icelandic, Welsh and Sámi language technology programmes.
- Managed funded projects end to end — award letters, instalment schedules, interim reporting, closure — across software localisation, lexicography and mobile keyboard development.
- Acted as technical counterpart between the foundation and its grantees, and argued open-licence conditions into grant terms so funded work could be built upon afterwards.

Independent Machine Translation Work

Karlsruhe, Germany · 2024

- Produced large-scale Western Armenian translation sets using the machine translation system built for the thesis, in collaboration with a diaspora language-revitalisation project, and extended and corrected the underlying parallel corpus from the output.

Student Research Assistant — Karlsruhe Institute of Technology

AI for Language Technologies Lab (AI4LT), Institute IAR · March 2023 – August 2023

- Built a streaming bot that joins virtual meetings automatically and feeds audio to the institute's Lecture Translator system for simultaneous translation.

Student Research Assistant — Karlsruhe Institute of Technology

Critical Information Infrastructures, Institute AIFB · November 2021 – August 2023

- Implemented and tested modules for a blockchain-based academic publication platform (Truffle, Ganache).
- Built a configurable cover-page generator integrated into the thesis submission pipeline, and automated recurring research tasks including citation search and tabular extraction.

WESTERN ARMENIAN LANGUAGE TECHNOLOGY

First neural machine translation system for Western Armenian

Master's thesis, KIT, with Prof. Jan Niehues · published at SIGUL, LREC-COLING 2024

- Built the first parallel corpus for the Western Armenian–English pair: approximately 147,000 sentence pairs collected from web sources and printed books across Germany, Turkey and Armenia, of which 52,879 are openly released.
- Fine-tuned Meta's NLLB, previously exposed only to Eastern Armenian, and measured cross-variant transfer; published held-out-domain test sets and a negative result on custom segmentation.
- Paper in the ACL Anthology (aclanthology.org/2024.sigul-1.42); corpus on GitHub; model served publicly with user feedback collection.

Ongoing open language infrastructure

2024 – present

- **hyw-corpus**: corpus expansion and normalisation, frequency lexicon, morphological data, spellchecking and locale data, with a documented source manifest and reproducible pipeline.
- **Speech recognition** for Western Armenian, fine-tuned on the limited aligned audio available.
- **Dictionary browser extension** for Western Armenian, published on the Chrome Web Store and Firefox Add-ons.

EDUCATION

M.Sc. Computer Science — Karlsruhe Institute of Technology

October 2020 – November 2023 · specialisation profile in Artificial Intelligence

- Thesis: Neural Machine Translation between the Endangered Language Western Armenian and English.

B.Sc. Computer Science — Karlsruhe Institute of Technology

October 2015 – September 2020

- Thesis: a longitudinal study of changes in smartphone usage in Germany and Turkey during the 2020 social-distancing period.

German Abitur — Deutsche Schule Istanbul

Istanbul, Turkey · 2015 · one of two schools in Turkey awarding the German Abitur

SELECTED PERSONAL ENGINEERING

- **Self-hosted LLM stack**: llama.cpp and vLLM serving quantised Qwen models on a single RTX 4090, behind a LiteLLM proxy, with Open WebUI, local search over MCP and a notification relay. Used daily as working infrastructure.
- **Home server**: Ubuntu, Docker, Tailscale, Pi-hole, running continuously and hosting the projects below.
- **Health tracking application**: React, Flask and SQLite, built for a specific dietary protocol.
- **Transit departure board**: Raspberry Pi and Flask, live local departures.
- **narinmatbaasi.com**: Next.js and Tailwind site for a historic letterpress print shop in Istanbul, with Armenian typography as its visual identity.

SKILLS

Programming	Python, SQL, JavaScript, TypeScript, Java, C, Haskell, Solidity, Bash
Data engineering	Polars, pandas, Dremio, Apache Iceberg, Parquet, PostgreSQL, MongoDB, lakehouse and medallion modelling, incremental and delta loading, REST and OAuth integrations, rate-limit and pagination handling, data lineage and reconciliation
Finance domain	Group consolidation, intercompany elimination, deferred revenue, ARR and MRR, budget versus actuals, cash-flow reporting, liquidity planning, accounting ledger and ERP data models
ML and NLP	PyTorch, Hugging Face Transformers, Fairseq, PEFT and LoRA fine-tuning, NLLB and Whisper, corpus construction, tokenisation and normalisation, evaluation design (BLEU, chrF, WER), OCR and computer vision (Tesseract, YOLO)
LLM tooling	llama.cpp, vLLM, Ollama, LiteLLM, GGUF quantisation, Model Context Protocol, agent and skill design, retrieval pipelines, prompt and evaluation harnesses
Cloud and DevOps	Azure (Blob Storage, Container Instances, Logic Apps, Key Vault and managed identity, App Service), Docker, Git, GitHub Actions, Linux administration, self-hosted infrastructure
Web	Django, Flask, REST API design, React, Next.js, Tailwind
Ways of working	Reverse engineering undocumented systems, technical writing and documentation, high-volume code review, direct client communication with non-technical stakeholders, Jira and trunk-based delivery
Languages	Western Armenian (native), Turkish (native), English (C2), German (C2), Italian (A2)