Kamyar Zeinalipour

AI engineer and NLP researcher: language models, instruction data and evaluation

k4myar@gmail.com kamyarzeinalipour.github.io linkedin.com/in/kamyarzeinalipour github.com/KamyarZeinalipour
huggingface.co/Kamyar-zeinalipour

I build and fine-tune language models and the data they learn from. At Villanova.AI I built the 1.71-million-conversation instruction dataset for its open European models; at YUKAI! I take fine-tuned models, retrieval pipelines and agents to production. Ph.D. in artificial intelligence from the University of Siena (2026, eccellente), with 31 papers, 19 as first author, at venues including ACL, EACL, LREC-COLING and IJCNLP-AACL.

Experience

Aug 2026 to now
Postdoctoral researcher, University of Siena, DIISM
- Language models for education and for languages with few resources. Published 17 open releases: each paper with its code, data and LoRA adapters, reproducible on a single GPU.

Nov 2025 to now
AI engineer, Villanova.AI, Cagliari
- Built Villanova-SFT-2603, the instruction-tuning data for Villanova-2B-2603: 1,711,114 conversations in five languages, curated, deduplicated and language-checked, released under Apache 2.0.
- Part of the team behind Villanova2, a dense 2.4B model and a 15.5B mixture of experts (3.9B active, 131k context) with open weights and data; co-author of Multi-SFT, the instruction data all four models cite.

Nov 2025 to now
Generative AI engineer, YUKAI!, Siena
- Fine-tuned models, retrieval pipelines and agentic systems in production, for domain-specific assistants and content generation in educational and creative applications.
- Built the agent that won Best Fooling AI Agent at Turing Hotel Italy (CLiC-it 2026): the one most often mistaken for a human.

Sep 2024 to Jun 2025
Visiting researcher, UMass Amherst, Inria Nice, University of Gothenburg
- Built a 22,000-pair human-verified dataset for detoxification that keeps the original meaning (UMass Amherst), the Detoxification Fidelity Index and a reinforcement learning pipeline for it (Inria), and a multilingual LLM study of plural nouns published at EACL 2026 (Gothenburg).

Apr 2022 to May 2023
Machine learning specialist, University of Siena
- Graph neural networks for ECG classification, from signals turned into graphs; published at WIRN 2022.

Jan 2022 to Mar 2023
Data scientist, Baker Hughes, with the University of Siena
- Models predicting the carbon monoxide and nitrogen oxide output of gas turbines, served as a REST API handling over a million predictions a day; cut tuning time by 20%. Two papers.

Projects

- **Tildea**: a product I run: it writes a resume for a job posting from the user's own profile and removes every claim they cannot back up.
- **Educational crossword generator**: turns course material into crosswords in Italian, Arabic, Turkish and English, with fine-tuned clue models.
- **ts2g**: a Python package that turns time series into graphs for graph neural networks.

Skills

Language models	fine-tuning (LoRA, QLoRA, PEFT), RLHF, instruction-data curation, retrieval-augmented generation and graph RAG, agents, evaluation
Frameworks	PyTorch, Hugging Face Transformers and PEFT, vLLM, DeepSpeed, LangChain, scikit-learn, XGBoost
Infrastructure	Linux, Docker, Git, CUDA, Slurm, Weights & Biases, Azure ML, Google Cloud
Programming	Python, SQL, R, Bash

Education

Nov 2022 to Feb 2026
Ph.D. in Information Engineering and Science, University of Siena
Thesis: From Language to Learning, an LLM-driven framework for multilingual educational content. Grade eccellente.

2017 to 2021
M.Sc. in Applied Mathematics (Data Science), University of Siena

Sep 2013 to Jan 2017
B.Sc. in Applied Statistics, University of Hormozgan, Iran

Teaching and languages

- Lecturer, Large Language Models for Developers, Zucchetti, July 2026: 20 hours, a lab for every lecture. Teaching assistant, Machine Learning, University of Siena, 2022 to 2026; supervised ten master's theses.
- Farsi (native), English (C1), Italian (B1).

31 papers, 19 as first author, and 17 open releases: kamyarzeinalipour.github.io/publications.html