Negar Foroutan-Eghlidi

Contact Info

Address: EPFL IC, BC 132 (Bâtiment BC), Station 14, 1015 Lausanne, Switzerland
Email: negar.foroutan@epfl.ch
Website: http://negar.foroutan.info

Interests

My research interests broadly encompass natural language processing (NLP) and machine learning, with a particular focus on improving the multilingual capabilities of large language models (LLMs), especially in low-resource settings. I work across the entire pipeline of training multilingual LLMs, including constructing high-quality pretraining datasets (language identification, filtering, and preprocessing), designing effective multilingual data mixtures, developing language-aware tokenization and architectures, and curating robust multilingual evaluation datasets.

Education

Current – 2019 | Ph.D. in Computer & Communication Sciences
Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland
Doctoral Research Assistant at NLP and LSIR labs under supervision of Antoine Bosselut and Karl Aberer
Thesis: Scaling Multilinguality: Addressing Low-Resource Language Limitations in Large Language Models

2013-2016 | M.Sc. in Computer Engineering (Artificial Intelligence)
Shiraz University, Shiraz, Iran
Thesis: Inferring Social Network Structure
Advisor: Dr. Ali Hamzeh

2009-2013 | B.Sc. in Computer Engineering (Software Engineering)
Shiraz University, Shiraz, Iran

Work Experience

Current – Jan. 2025 | Research Intern - Google Research, Zurich, Switzerland
Working on a project to optimize long-context inference, improving LLMs' efficiency in processing and understanding extended inputs.

Excpected 2019-2025 | Doctoral Research Assistant - NLP & LSIR, EPFL, Lausanne, Switzerland
Contributed to multiple research projects on multilingual LLMs, covering the entire training pipeline. Led the multilingual effort within the SwissAI initiative. Collaborated on projects with Google, Cohere, and HuggingFace. Supervised junior researchers and summer interns. Served as a teaching assistant in several courses.

2018-2019 | Research Assistant - Machine Learning and Optimization Laboratory, EPFL, Switzerland
I was involved in the mlbench project, a benchmark framework for distributed machine learning.

Summer 2017 | Research Intern - Data Analytics Laboratory, ETH, Zurich, Switzerland
As an intern in Thomas Hofmann's lab working under the supervision of Carsten Eickhoff, I worked on a modular, patient-centric information retrieval system designed for precision oncology applications. The result of the project was a submission to the TREC 2017 Precision Medicine track.

Spring 2017 | Research Intern - Max Planck Institute for Software Systems, Kaiserslautern, Germany
As an intern under the supervision of Manuel Gomez Rodriguez, I worked on a project analyzing the dynamics of citation networks: quantifying the value of a set of published papers and modeling knowledge diffusion across a citation network.

2016-2017 | R&D Engineer - Center of Intelligent Vision & Image Processing, Shiraz University, Shiraz, Iran
I worked on projects focused on object detection, facial expression analysis, and real-time face recognition and tracking.

Technical Skills

Programming: Python, Java, Scala, C/C++, MATLAB
Framework & Libraries: PyTorch, Jax, TensorFlow, Spark, OpenCV
Miscellaneous: Git, Docker, LaTeX, Shell Scripting
Operating Systems: macOS, Linux, Windows

Languages

Persian: Native Proficiency
English: Full Professional Proficiency
French: Elementary Proficiency

Publications

1. **N. Foroutan**, Jakhongir Saydaliev, Ye Eun Kim, Antoine Bosselut, "Supervised Contrastive Learning for Low-Resource Language Identification" *arXiv 2025*.
2. A. Romaneo, **N. Foroutan**, Anna Sotnikova, et al. "INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge." *ICLR 2025 (Spotlight)*.
3. B. Borges*, **N. Foroutan***, D. Bayazit*, et al., "Could ChatGPT get an Engineering Degree? Evaluating Higher Education Vulnerability to AI Assistants." *PNAS 2024*.
4. C. Fierro, **N. Foroutan**, D. Elliott, and A. Søgaard, "How Do Multilingual Models Remember? Investigating Multilingual Factual Recall Mechanisms." *arXiv 2024*.
5. D. Bayazit, **N. Foroutan**, Z. Chen, G. Weiss, and A. Bosselut, "Discovering Knowledge-Critical Subnetworks in Pretrained Language Models." *EMNLP 2024*.
6. **N. Foroutan**, M. Banaei, K. Aberer, and A. Bosselut, "Breaking the Language Barrier: Improving Cross-Lingual Reasoning with Structured Self-Attention." *EMNLP 2023 - Findings*.
7. Y. Karoui, R. Lebret, **N. Foroutan**, and K. Aberer, "Stop Pre-Training: Adapt Visual-Language Models to Unseen Languages." *ACL 2023*.
8. **N. Foroutan**, M. Banaei, R. Lebret, A. Bosselut, and K. Aberer, "Discovering Language-neutral Sub-networks in Multilingual Language Models." *EMNLP 2022*.
9. **N. Foroutan**, A. Romanou, S. Massonnet, R. Lebret, and K. Aberer, "Multilingual Text Summarization on Financial Documents." *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC, 2022*.
10. **N. Foroutan Eghlidi**, and M. Jaggi, "Sparse Communication for Training Deep Networks." *Workshop on "Beyond first-order methods in ML systems" at ICML 2020*, arXiv preprint arXiv:2009.09271 (2020).
11. **N. Foroutan** and A. Hamzeh, "Discovering the Hidden Structure of a Social Network: A Semi-Supervised Approach." *IEEE Transactions on Computational Social Systems*, pp. 14-25. 2017.
12. **N. Foroutan Eghlidi**, J. Griner, N. Mesot, L. von Werra and C. Eickhoff, "ETH Zurich at TREC Precision Medicine 2017." *Proceedings of the 26th Text Retrieval Conference (TREC)*, 2017.

Selected Projects

2024-2025 | Supervised Contrastive Learning for Low-Resource Language

Language identification (LID) is key to curating multilingual LLM pretraining corpora, but low-resource languages–often limited to single-domain sources like the Bible–still underperform. To resolve these class imbalance and bias issues, we propose a supervised contrastive learning (SCL) approach to learn domain-invariant representations. Our analysis shows it improves LID performance on out-of-domain data by up to 3.7%, demonstrating its effectiveness in enhancing LID models.

2024-2025 | Language-aware Tokenization

Current LLMs often underperform in non-target languages, especially low-resource ones, partly due to tokenization disparities before training even begins. Inconsistent tokenization across languages leads to unequal treatment, disadvantaging certain language communities. This has real-world consequences, including higher costs, longer processing times, and reduced contextual capacity. To address this, we propose a parity-aware BPE to create a more equitable tokenizer for multilingual models, ensuring fairer performance and accessibility across languages.

2024-2025 | Multilingual Data Mixture

Multilingual LLMs are trained across many languages, but data availability varies widely, making this a multitask learning problem with severe data imbalance. English, the highest-resource language, often dominates, complicating the challenge of balancing languages during pretraining—especially given its high costs. This project explores different data mixture strategies, including heuristic and proxy-based multilingual sampling. We also examine the role of a pivot language (the one with the most data) and the effects of curriculum learning, where languages are introduced gradually during training.

2024 | FineWeb-2: Multilingual Pretraining Dataset

FineWeb-2 is a high-quality multilingual pretraining dataset covering over 1,000 languages. Built from 96 CommonCrawl snapshots (2013–2024) and processed with the datatrove library, it includes approximately 8 TiBs of compressed text and nearly three trillion words. The dataset is carefully deduplicated and filtered, providing curated data for 1,893 unique language-script pairs. This project was developed in collaboration with Hugging Face.

2018-2019 | mlbench: Distributed Machine Learning Benchmark

This project aimed to create a reference collection of distributed machine learning algorithm implementations across various frameworks and system platforms. My focus was on supervised learning, including deep learning and linear models. We defined standardized tasks and datasets to enable fair and precise comparisons of algorithms, frameworks, and hardware.

Teaching Experience

Spring 2024 | Modern NLP - EPFL
Guest lecturer; also designed exercises, programming assignments, project, and exams; graded assignments and exams.

Fall 2022/23/24 Spring 2022 | Distributed Information Systems - EPFL
Lead weekly lab sessions and designed assignments and projects. Graded exams and projects.

Spring 2021 | Database systems - EPFL
Graded assignments, projects and exams; held lab sessions and gave guidance to students for their projects.

Fall 2021 Fall 2020 | Advanced Information, Computation, Communication I - EPFL
Graded programming assignments and exams; led weekly lab sessions and gave guidance to students.

Fall 2015 | Statistical Pattern Recognition - Shiraz University
Graded programming assignments and projects.

Spring 2015 | Machine Learning - Shiraz University
Graded programming assignments and projects; led weekly lab sessions and gave guidance to students.

Spring 2014 | Artificial Intelligence - Shiraz University
Led weekly lab sessions; gave guidance to students and graded their assignments and projects.

Fall 2012 | Fundamentals of Computer and Programming Using Python - Shiraz University
Constructed the syllabus and prepared the course material (programming assignments, labs, and projects); led weekly lab sessions; guided the students and graded their assignments and projects.

Spring 2012 | Principles of Programming Using C - Shiraz University
Prepared and graded programming assignments and projects; led weekly lab sessions and gave guidance to students.

Fall 2014 Fall 2011 | Advanced Programming Using Java - Shiraz University
Prepared and graded programming assignments and projects; led weekly lab sessions and gave guidance to students.

Professional & Extracurricular Services

2022-2025 | PC Member & Reviewer
ACL 2024, ACL 2025, NAACL 2025, ICLR 2025, EMNLP 2022, EMNLP 2023, ARR June 2024, RepNLP 2023, BiAlign 2025

2016 | AIESEC Team Member, Shiraz, Iran.

Sep. 2016 | Organization team member of TEDxShirazUniversity 2016.

2009 - 2015 | Students' Scientific Council (SSC), CSE Department, Shiraz University
SSC is the student committee concerned with directing the department extra-curriculum activities. I had the chance to be SSC's chair from 2011 to 2012.

Dec. 2014 | Co-organizer of an Hour of Code Event
Organized a one-day workshop, participating in the Hour of Code program, for tens of high-school and middle-school students and taught them the basics of computer programming and algorithmic thinking.

2010-2014 | BreakTime In University (BTiU)
BTiU is a three-day annual conference consists of tens of parallel workshops held by a group of university students during the summer at Shiraz University. Hundreds of talented high-schoolers attend this event to learn more about various study majors, practice teamwork, life and social skills, and learn how to be creative and innovative. I had the chance to be a part of the organizing team for five years.

May 2012 | Member of Conference Organizing Committee
Internet and technical services assistant at the 16^{th} CSI International Symposiums on Computer Architecture & Digital Systems (CADS 2012) and Artificial Intelligence & Signal Processing (AISP 2012) held at Shiraz University, Shiraz, Iran.

Last Update: February 2025