Amirreza Sattarzadeh Khanehbargh

Istanbul, Türkiye · amirreza.sattarzadeh@bogazici.edu.tr
amirreza-stz.github.io · Google Scholar · GitHub · LinkedIn **Seeking PhD positions starting in autumn 2027**

Research interests: Natural language processing (NLP) and machine learning (ML), centred on language models (LMs): tokenization and input representations, efficient adaptation and fine-tuning, and interpretability of what models encode. **Bringing NLP and ML to other fields:** biology (protein language models and BioNLP), the social sciences (hate speech in multilingual text), finance (market-derived sentiment), and low-resource, less-studied languages and their linguistic structure (e.g. Universal Dependencies; I have worked with Turkish UD data).

Research Experience

Graduate Research Assistant, LifeLU (ERC Consolidator Grant) Oct 2024 – present
TabiLab, Boğaziçi University · PI: Prof. Arzucan Özgür · Understanding the Language of Life: Identifying and Characterizing the Language Units in Protein Sequences

- Study how the choice of tokenization (amino acid, subword and evolution-aware units) interacts with fine-tuning and parameter-efficient adaptation in protein language models (PLMs), including ESM-2 models pretrained from scratch with a teammate.
- Built a benchmark of **6,500+ training runs** on **14 protein function, localization and stability tasks**. Evolution-aware units outperformed frequency-based subwords in several settings and matched or surpassed amino acid tokens at some vocabulary sizes.
- Tested adding predicted protein dynamics features to PLM representations. Next: interpreting the learned units, other knowledge fusion methods and their interaction with tokenization, dynamic and task-specific tokenization.
- Regular presenter in the team's weekly reading club.

Developer and Researcher, PARI project Oct 2026 – present
Hrant Dink Foundation with Boğaziçi and Sabancı universities, Istanbul
Funded by the EU and the Friedrich Naumann Foundation

- Recently joined as the project's student researcher, working with its faculty, postdoctoral and PhD researchers on an open-source NLP tool for detecting hate speech and discriminatory content in Turkish and Arabic.

Team member, ELSA Health Privacy Challenge, CAMDA 2026 at ISMB 2026 Jul 2026
Organized by EMBL, CISPA Helmholtz Center and the University of Helsinki within the EU ELSA project

- Equal-contribution first author with Buse Giledereli and Başar Temiz. Benchmarked single-cell RNA-seq generators, including an LLM-based one (Cell2Sentence), for biological utility against membership inference privacy, and showed that how non-highly variable genes are reconstructed is the main privacy lever.
- Our work was selected for a talk and presented as a talk and a poster at ISMB 2026, Washington, DC.

Research Assistant Jul 2023 – Sep 2024
Iran University of Science and Technology (IUST), Tehran · Supervisor: Dr. Reza Entezari-Maleki

- *Financial market trend analysis using FinBERT and sentiment analysis in online text data* (Feb – Sep 2024): labelled posts by the market move that follows them instead of human sentiment labels and fine-tuned domain language models on these labels (up to +11% trend prediction accuracy; 89.6% with market and temporal context in the prompt); extended into the arXiv preprint.
- *Deep learning and neural networks for algorithmic trading* (Jul 2023 – Feb 2024).

Publications and Posters

- **Sattarzadeh Khanehbargh, A.**, Suyunu, B., Dolu, Ö., and Özgür, A. (2026). *Tokenization-Aware Protein Language Modeling with Evolution-Guided Units and Biological Knowledge Fusion.* 25th European Conference on Computational Biology (**ECCB 2026**), Geneva. Poster C-P.51, presented in person (poster page).
- **Sattarzadeh Khanehbargh, A.** and Özgür, A. (2026). *From Protein Sequences to Biological Language Units: Frequency- and Evolution-Aware Tokenization for Protein Language Models.* Lectures on Computational Linguistics (**LCS 2026**), Bolzano, Italy. Accepted poster.
- Giledereli, B.*, Temiz, B.*, and **Sattarzadeh Khanehbargh, A.*** (2026). *Benchmarking Recent Single-Cell Synthetic Data Generation Methods on the OneK1K Dataset.* ELSA Health Privacy Challenge, **CAMDA 2026 at ISMB 2026**, Washington, DC (selected talk and poster). Preprint, arXiv:2610.31454. *Equal contribution.
- Moradi-Kamali, H., Rajabi-Ghozlou, M.-H., Ghazavi, M., Soltani, A., **Sattarzadeh, A.**, and Entezari-Maleki, R. (2025). *Market-Derived Financial Sentiment Analysis: Context-Aware Language Models for Crypto Forecasting.* Preprint, arXiv:2502.14897. 12 citations on Google Scholar, including its earlier version.

Education

M.Sc. in Computer Engineering Oct 2024 – expected Jul 2027
Boğaziçi University, Istanbul, Türkiye (English-taught)

- Supervisor: Prof. Arzucan Özgür, Text Analytics and Bioinformatics Lab (TabiLab) · GPA 3.25 / 4.0
- Thesis (in progress): *From Protein Sequences to Biological Language Units: Frequency- and Evolution-Aware Tokenization for Protein Language Models*

B.Sc. in Computer Engineering Sep 2018 – Sep 2024
Iran University of Science and Technology (IUST), Tehran, Iran

- Final project: *Financial Market Trend Analysis using Sentiment Polarity in Online Text Data* (Dr. Reza Entezari-Maleki), extended into the arXiv preprint.

High school diploma, Mathematics and Physics 2014 – 2018
SAMPAD (NODET: National Organization for Development of Exceptional Talents), selective gifted school, Bonab, Iran · ranked 1st among the school's Mathematics and Physics students

Industry Experience

Python Developer (Backend and Data), Roshan-AI Jun 2022 – Nov 2023
Tehran · deep learning company · joined through a competitive backend internship, then part-time developer

- *Harf* (Persian speech recognition): built the data pipeline for crawling, cleaning and preprocessing speech–text data, growing the training data in size and quality, which improved the speech models; Django APIs for deployment.
- *Replai* (NLP comment and reply classification): Django backend.

Junior Software Engineer (freelance) Sep 2019 – Feb 2022
Bonab, Iran

- Telegram bot that crawls Iranian stock market data and screens growth stocks.
- Django backend for a charity website, in a team of five.

Teaching and Mentoring

- **Teaching Assistant, IUST** (Sep 2023 – Jul 2024), seven course sections: Head TA for Data Science; Software Engineering Mentor and Scrum Master for Systems Design and Analysis; Scrum Master for Software Engineering II; TA for Advanced Programming (two sections), Fundamentals of Computer Programming, Computer Workshop.
- **Python Instructor**, independent, online (Jan 2023 – Jun 2024): introductory Python for beginners.
- **Volunteer Instructor**, Pooyandegan-e-Daneshpazhooh-e-Mehr (BonabNGO) (Summer 2023): designed and ran a free online Python and computer engineering fundamentals course for students in the region.

Honors, Service and Skills

Honors	Graduate entrance exam, Iran (2024): 764th in Computer Engineering and 398th in Information Technology nationwide; chose Boğaziçi instead University entrance exam (2018): 233rd of ~160,000 (Mathematics), 930th of ~60,000 (Language) National Mathematical Olympiad, 2nd stage (2016); Paya Scientific League, 2nd stage in Mathematics (2015) and Biology (2016)
Tests	GRE Quantitative Reasoning 165 / 170 (2024)
Service	Board member (since 2021), Pooyandegan-e-Daneshpazhooh-e-Mehr (BonabNGO), an educational NGO; core team, Iran Innovate (Iran branch of Innovate Decentral), 2023 – 2024
Societies	ISCB; AILC (Italian Association for Computational Linguistics)
Tools	Python, PyTorch, Hugging Face, Weights & Biases, Linux and remote GPU servers, Git, Django
Languages	Persian, Azerbaijani and Turkish (native or bilingual); English (full professional); Arabic (elementary)

Referees

Further references available on request.

- Prof. Arzucan Özgür, Boğaziçi University (Dept. Chair); M.Sc. supervisor, LifeLU PI · arzucan.ozgur@bogazici.edu.tr
- Dr. Reza Entezari-Maleki, former Head of Software Engineering, IUST; B.Sc. supervisor, coauthor · entezari@iust.ac.ir
- Dr. Mehrdad Ashtiani, IUST; supervised my mentoring, Scrum Master and TA roles · m_ashtiani@iust.ac.ir
- Dr. Hassan Naderi, IUST; supervised my Head TA role in Data Science · naderi@iust.ac.ir
- Dr. Maryam Khodadadi, University College London (UCL); my CAD course instructor at IUST · m.khodadadi@ucl.ac.uk
- Dr. Reza Yousefi Maragheh, Walmart Global Tech; mentor, fellow BonabNGO board member · ryousefimaragheh@acm.org
- Dr. Mohsen Mohammadi, Roshan-AI; my direct manager (Harf speech recognition) · mohammadi.mohsen@gmail.com