Hossein Rouhizadeh

PhD Candidate in Computer Science | AI & NLP Researcher

Email | LinkedIn | GitHub | Google Scholar

Research Statement

Computer science PhD candidate specializing in natural language processing and large language models (LLMs), with a proven track record of developing production-ready AI solutions. Published research in prestigious venues including Nature Scientific Data, demonstrating expertise in scaling ML systems across multiple languages and domains. Proven ability to lead technical projects and mentor junior researchers, complemented by hands-on experience with modern AI technologies and deployment at scale. Developed state-of-the-art language models achieving top performance in international competitions (1st place, #SMM4H 2023).

Research Interests

Machine Learning for Healthcare Applications; Natural Language Processing; Large Language Models (LLMs); Information Extraction; Information Retrieval; lexical Semantics

Education

- **University of Geneva** *June. 2022 - Present*
  PhD Candidate in Computer Science Geneva, Switzerland
  - Thesis: "Foundation Models for Information Extraction from Electronic Health Records"
  - Key Achievements:
    - Developed BioWiC, a novel benchmark dataset for evaluating biomedical concept representation in LLMs
    - Created multilingual biomedical concept normalization system supporting 5 European languages
    - Led research on detecting human vs machine-generated biomedical case reports using LLMs
- **Shahid Beheshti University** *Sep. 2016 - Sep. 2019*
  Master of Science in Computer Engineering Tehran, Iran
  - Thesis: "Multilingual Word Sense Disambiguation Using Deep Learning Methods"
  - Key Achievements:
    - Developed state-of-the-art multilingual Word Sense Disambiguation system
    - Created first Persian sense-annotated corpus (PerSemCor)
    - Built standard test set for Persian all-words WSD

Work and Research Experience

- **Data Science for Digital Health (DS4DH) group** *June. 2022 - present*
  Doctroal Research Assistant - AI/NLP Engineer Geneva, Switzerland
  Project: Natural Language Understanding for Electronic Health Records (NLU4EHR)
  - Fine-tuned for generative LLMs (e.g, Llama-3, and Mistral) for concept normalization task
  - Fine-tuned discriminative LLMs using SBERT
  - Leveraged generative LLMs as encoders using LLM2VEc framework
  - Integrated multilingual biomedical entity retrieval using Elasticsearch
  - Developed innovative RAG-based framework enhancing contextual representations of medical terminology
  - Implemented ranking fusion techniques to ensemble results of different LLMs
- **Dadmatech** *Oct. 2020 - May 2022*
  Research Scientist Tehran, Iran
  Knowledge Extraction from Texts
  - Fine-tuned discriminative LLMs for entity typing and extraction
  - Developed end-to-end pipeline for knowledge extraction from free texts
  - Implemented co-reference resolution and relation classification systems
  - Deployed models via Flask web services

- **NLP Lab, Shahid Beheshti University** *Mar. 2020 - Aug 2020*
  ML/NLP Engineer Tehran, Iran
  Text Processing on Persian COVID-19 Tweets
  - Fine-tuned discriminative LLMs for Persian topic classification and sentiment analysis
  - Implemented LDA for tweet clustering
  - Trained specialized word embedding models (Word2Vec, GloVe, fastText) on COVID-19 tweets

Publications

- Journal Articles
  - **Hossein Rouhizadeh**, et al. (2024). "A Dataset for Evaluating Contextualized Representation of Biomedical Concepts in Language Models", Nature Scientific Data
  - **Hossein Rouhizadeh**, et al. (2022). "Persian-WSD-Corpus: A Sense Annotated Corpus for Persian All-words Word Sense Disambiguation", International Journal of Web Research
  - **Hossein Rouhizadeh**, et al. (2024). "Comparative Analysis of Large Language Models for Multilingual Biomedical Concept Normalization", Under review, npj Digital Medicine
  - Anthony Yazdani, **Hossein Rouhizadeh**, et al. (2024). "CONORM: Context-Aware Entity Normalization for Adverse Drug Event Detection", Under review, npj Digital Medicine
  - Anthony Yazdani, [...], **Hossein Rouhizadeh**, et al. (2024)., "CT-ADE: An Evaluation Benchmark for Adverse Drug Event Prediction from Clinical Trial Results", Under review, Nature Scientific Data
  - Boya Zhang, [...], **Hossein Rouhizadeh**, et al. (2024). "A Dataset for Evaluating Clinical Research Claims in Large Language Models", Under review, Nature Scientific Data
- Conferences & Workshops Articles
  - **Hossein Rouhizadeh**, et al. (2025). "Exploring Zero-Shot Cross-Lingual Biomedical Concept Normalization via Large Language Models", Medical Informatics Europe
  - Anthony Yazdani, **Hossein Rouhizadeh**, et al. (2025). "Leveraging Large Language Models for Synthetic Data Generation to Enhance Adverse Drug Event Detection in Tweets", Medical Informatics Europe
  - Anthony Yazdani, **Hossein Rouhizadeh**, et al. (2023). "DS4DH at #SMM4H 2023: Zero-Shot Adverse Drug Events Normalization using Sentence Transformers and Reciprocal-Rank Fusion", #SMM4H Workshop
  - **Hossein Rouhizadeh**, Douglas Teodoro (2022). "DS4DH at SemEval-2022 Task 11: Multilingual Named Entity Recognition Using an Ensemble of Transformer-based Language Models", SemEval Workshop
  - **Hossein Rouhizadeh**, et al. (2021) "Persian SemCor: A Bag of Words Sense-Annotated Corpus for the Persian Language", Global WordNet Conference
  - **Hossein Rouhizadeh**, et al. (2020) "Word Sense Disambiguation with Distributional Semantic Expansion for the Persian Language", International Conference on Computer and Knowledge Engineering

Teaching & Mentoring Experience

- University of Geneva (2022-Present)
  - Teaching Assistant
    - Applied Machine Learning in Healthcare: ChatGPT and Beyond
  - Mentoring:
    - 2 Medical Doctors (Master's students at University of Geneva)
    - 2 Computer Science Master's students (University of Lyon 1)
- Shahid Beheshti University (2016-2019)
  - Teaching Assistant:
    - Natural Language Processing, Deep Learning, Artificial Intelligence, Ontology Construction

Technical Skills

- **Deep Learning Tools:** PyTorch, PyTorch-Lightning, Keras, TensorFlow, Hugging Face
- **LLM Technologies:** LlamaIndex, LLM2VEc, RAG, LLM fine-tuning pipelines, Sentence-BERT
- **Vector Databases:** Qdrant, Chromadb
- **NLP Tools:** SpaCy, Stanford CoreNLP, Flair, pySBD, Gensim, NLTK
- **Visualization Tools:** Matplotlib, Pyqtgraph, Seaborn
- **Web Development Tools:** Flask, RESTful APIs
- **Development Tools:** Git, Docker

Honors and Awards

- First Place: SMM4H 2023 Workshop Challenge - Adverse Drug Event Mention Normalization

Languages

- English (Fluent)