KTH Royal Institute of Technology
Stockholm, Sweden
📱 *+46 (73) 49 13 530*
✉ *amlk@kth.se*
🌐 *amirlayegh.github.io*
in *amirhosseinlayegh*

Amirhossein Layegh
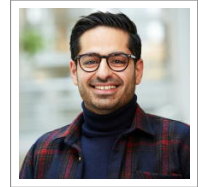
Research Interests

Scalable Machine Learning (ML) and Deep Learning (DL), Information Retrieval, RAG Systems, Large Language Models, Natural Language Processing, Model Optimization and Fine-tuning

---

Education

2021 – now   **PhD in Information and Communication Technology**, *KTH Royal Institute of Technology*, Sweden
Thesis Title: Advancing Information Extraction through Large Language Models
Advisers: Prof. Mihhail Matskin and Dr. Amir H. Payberah
Research focuses on developing novel retrieval and information extraction systems using LLMs, with emphasis on:
- Designing prompting strategies that leverage external knowledge bases for zero-shot scenarios
- Developing multi-stage refinement approaches for improved extraction accuracy
- Creating fairness-aware data augmentation techniques using knowledge graphs
- Contributing to medical information retrieval systems for healthcare applications

2018 – 2019   **MSc in Big Data Science**, *Queen Mary University of London (QMUL)*, UK
Thesis Title: Implementation of a Recommendation System for a retail store based on basket analysis.
Adviser: Dr. Arman Khouzani
Developed a recommendation system for Koolbitz, a London-based retailer, using the Apriori algorithm for basket analysis. The system identifies patterns in purchase data to suggest products to customers, enhancing sales and customer experience.

2012 – 2017   **BSc in Software Engineering**, *Ferdowsi University of Mashhad (FUM)*, Iran
Adviser: Dr. Mohsen Kahani
Completed a final project on developing a database of academic publications for current and former Ferdowsi University affiliates. Utilized web scraping techniques to automate publication data collection, significantly enriching the university's academic resources.

Industrial Experience

2020 – 2021   **Data Scientist**, *Neshan Maps*, Iran
Employed advanced speed estimate techniques on large-scale user datasets to analyze and determine traffic flow for individual road segments. Developed models to accurately forecast total traffic flow, enhancing the app's navigation accuracy and user satisfaction.

2019 – 2020   **Data Analyst**, *Koolbitz Ltd.*, UK
Developed and implemented machine learning algorithms to analyze large sales datasets, identifying key patterns and trends for product recommendations. This significantly improved targeted marketing strategies and customer engagement.

2017 – 2018   **Software Developer**, *Ranir*, Iran
Engaged as an Oracle Application Development Framework (ADF) developer, focusing on building Java-based enterprise applications.

Supervision

Supervision

- MSc thesis supervisor of Master students
  - **Deciding Agents Behaviours Through LLMs:** This thesis investigates using LLMs for decision-making and behavior control in AI agents within gaming virtual environments. It aims to enhance AI agent design by integrating natural language processing for intuitive human-AI interactions with potential gaming, virtual reality, and digital assistant applications.
  - **Enhancing Breast Cancer Information Retrieval:** This project focuses on developing, evaluating, and optimizing Retrieval-Augmented Large Language Models specifically for improving breast cancer information retrieval accuracy and efficiency. It aims to contribute to the medical field by providing better tools for accessing and understanding breast cancer-related data.
  - **Designing Novel Effective Methods for Large Language Model Compression**: Guiding research on pruning and lightweight techniques for compressing large language models, specifically targeting 7B - 13B models. The objective is to reduce these models to fewer than 3B parameters while preserving their performance, employing strategies such as knowledge distillation through teacher-student models.

Publications

- REA: Refine-Estimate-Answer Prompting for Zero-Shot Relation Extraction, Presented as full paper at NLDB 2024, LINK
  Developed a novel multi-stage prompting framework for improved information extraction in zero-shot scenarios
- Wiki-based Prompts for Enhancing Relation Extraction using Language Models, Presented full paper at SAC 2024, LINK
  Created a knowledge-enhanced prompting approach leveraging external knowledge bases for improved extraction
- ContrastNER: Contrastive-based Prompt Tuning for Few-shot NER, presented as full paper at COMPSAC 2023, LINK
  Introduced a novel contrastive learning framework for few-shot entity extraction
- A Unified Framework for Evaluating Mathematical Reasoning in LLMs: New Metrics and Insights, ECIR 2025, under-review
  Proposed novel evaluation metrics for assessing LLM reasoning capabilities
- Dataclouddsl: Textual and Visual Presentation of Big Data Pipelines (2022), LINK
- A survey of big data pipeline orchestration tools from the perspective of the datacloud project (2021), LINK

Skills

- **Programming Languages & Frameworks:** Python, PyTorch, Hugging Face Transformers, CUDA
- **Large Language Models:** Prompt Engineering, RAG, Parameter Efficient Fine-tuning (LoRA, Adapter), Fine-tuning, Zero-shot/Few-shot Learning, Contrastive Learning, Knowledge-Enhanced Prompting
- **MLOps & Infrastructure:** SageMaker, ZenML, Git, Docker
- **Distributed Systems:** Spark, Distributed Training, Parallel Computing
- **Deep Learning:** Neural Networks, Transformers, Foundation Models