

İrem Karaca

Tübingen, Germany | iremkaracacs@gmail.com | [iremkr.github.io](https://github.com/iremkr) | [linkedin.com/in/irem-karaca](https://www.linkedin.com/in/irem-karaca)

EDUCATION

Eberhard Karls University of Tübingen

M.Sc. in Machine Learning

Tübingen, Germany

Oct 2024 – Mar 2027 (expected)

- ELIZA research-oriented master's scholar.
- Relevant Coursework: AI Safety, Understanding Language Models, Natural Language Processing

Koç University, College of Engineering

B.Sc. in Computer Engineering

Istanbul, Türkiye

Sep 2019 – Jun 2024

- GPA: 3.75/4.00; graduated Summa Cum Laude.

RESEARCH EXPERIENCE

Research Assistant

ELLIS Institute Tübingen

Feb 2025 – Present

Tübingen, Germany

- Member of the AI Tutor project (supervised by Wieland Brendel), an adaptive tutoring system; conducted a research study on probabilistic student modeling for curriculum personalization (see Projects).

Research Intern

Brussels Human Robotics Research Center, Vrije Universiteit Brussel

Jul 2023 – Sep 2023

Brussels, Belgium

- Developed machine learning models to estimate muscle force for upper-limb exoskeletons using ultrasound measurements (supervised by Tom Verstraten).

Research Assistant

Intelligent User Interfaces Laboratory, Koç University

Nov 2019 – Nov 2020

Istanbul, Türkiye

- Assisted research on sketch-based interfaces and applications in creative processes (supervised by Metin Sezgin).

PROJECTS

Open-Source Replication of Subliminal Learning

Summer 2026

- Reimplemented subliminal learning experiments with open-source Qwen models in an offline cluster environment.
- Studied whether hidden behavioral traits can transfer through unrelated synthetic data, with implications for data poisoning and emergent misalignment.

Course Projects — AI Safety, University of Tübingen

Summer 2026

- **Emergent Misalignment:** Reproduced emergent misalignment experiments by fine-tuning Qwen3-4B-Instruct with LoRA on risky financial-advice data; analyzed the effect of data mix and LoRA rank on misalignment rates.
- **LLM Jailbreaking:** Tested manual jailbreak strategies and implemented a random-search suffix attack using token-level log-probabilities; evaluated attack success with automated safety judges and manual inspection.
- **Representation Engineering and Activation Steering:** Trained linear probes for harmfulness across model layers and applied activation steering / directional ablation to study refusal behavior in Qwen3-4B-Instruct.

Student Modeling for Adaptive Learning Systems

Winter 2025/26

- 9 ECTS graded research project at ELLIS Institute Tübingen, supervised by Dr. Wieland Brendel.
- Designed probabilistic student models based on Bayesian Knowledge Tracing variants with prerequisite-weighted learning dynamics and ran a 27-condition simulation study showing a robust efficiency–mastery trade-off between ZPDES variants.

FELLOWSHIPS

Technical AI Safety Course and Project Participant

BlueDot Impact

May 2026 – Present

Remote

- Completed a 30+ hour facilitated technical AI safety course covering core concepts in AI alignment.
- Working on an AI safety research project focused on Replication of Subliminal Learning (see Projects).

Master's Student Fellow <i>Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA)</i>	Oct 2024 – Present Tübingen, Germany
<ul style="list-style-type: none"> Awarded a scholarship and supervision through ELIZA, in collaboration with the European Laboratory for Learning and Intelligent Systems (ELLIS), as part of my master's program in machine learning. 	
Machine Learning Bootcamp Participant <i>Google Developers Group & inzva</i>	May 2023 – Oct 2023 Remote
<ul style="list-style-type: none"> Participated in a 5-month bootcamp including Coursera courses and weekly classes focused on core machine learning concepts and practical applications. 	
Young Women Leaders Fellow <i>Yenibirlider Association & Turkish Confederation of Employer Associations</i>	Jun 2021 – Apr 2022 Istanbul, Türkiye
<ul style="list-style-type: none"> Completed training and participated in seminars on leadership, career management, project management, social skills, case studies, and future job trends. 	

INDUSTRY EXPERIENCE

Computer Vision Engineer Intern <i>Delivers.AI</i>	Jun 2022 – Sep 2022 Istanbul, Türkiye
<ul style="list-style-type: none"> Worked on object detection for an autonomous delivery robot, improving the performance of a machine learning model using YOLOv4 and achieving a higher mAP compared to the previous model. 	
Software Development Intern <i>OBSS</i>	Jul 2021 – Aug 2021 Istanbul, Türkiye
<ul style="list-style-type: none"> Developed a full-stack online marketplace web application using Java, Spring Boot, React, and MySQL, enhancing user experience and functionality. 	

TEACHING EXPERIENCE

Teaching Assistant <i>Computer Systems & Programming (by Aykut Erdem), Koç University</i>	Feb 2022 – Feb 2023 Istanbul, Türkiye
<ul style="list-style-type: none"> Conducted office hours and lab sessions, providing academic support to students and assisting with programming exercises. 	
Section Leader <i>Code in Place, Stanford University</i>	Apr 2021 – May 2021 Remote
<ul style="list-style-type: none"> Conducted problem-solving sessions with 10 students from diverse backgrounds and ages, teaching introductory Python programming skills over a 5-week course. 	
Tutor <i>CS 101 & Girls Who Code Programs</i>	Feb 2021 – May 2021 Remote
<ul style="list-style-type: none"> Taught introductory Python to high school students, including 8 weeks of 75-minute lectures. 	

VOLUNTEERING

Mentorship Program Coordinator <i>inzva</i>	Aug 2022 – Aug 2024 Istanbul, Türkiye
<ul style="list-style-type: none"> Coordinated two annual cohorts of a mentorship program connecting Turkish computer science students with mentors from academia and industry; scaled the program to 12 mentors and 24 mentees per cohort. Introduced industry and academia talks, organized hybrid closing ceremonies for mentors abroad, and built a sustainable structure that continued after handoff. 	
Corporate Relations Manager <i>Koç University ACM Student Chapter</i>	Aug 2020 – Jan 2021 Istanbul, Türkiye
<ul style="list-style-type: none"> Managed a 15-person corporate relations team and coordinated technology and career events in collaboration with industry professionals. 	

SKILLS

Programming: Python, Java, C/C++, SQL, JavaScript
Tools/Frameworks: PyTorch, Git, Linux, SLURM/HPC clusters, HuggingFace, TensorFlow, L^AT_EX
Languages: Turkish (native), English (fluent), German (intermediate)