Serhan YILMAZ

✉ serhan.yilmaz@sabanciuniv.edu | 🔗 yilmaz-serhan | ⓖ serhanylmz | 🌐 serhanyilmaz.org

RESEARCH INTERESTS: **Agentic AI**, **Reasoning**, **Test-Time Scaling**, **Compound AI Systems**

EDUCATION

Computer Science — *Bachelor of Science*   SEP 2021 - JUN 2025
Sabanci University   cGPA: **3.59**/4
- ***Relevant Coursework:*** Machine Learning, NLP, Artificial Intelligence, Data Science, Image Processing, Linear Algebra, Statistics, Algorithms, Data Structures, Advanced Programming, Operating Systems, Databases
- Ranked **top 0.02%** among 2.9 million students in the National University Entrance Exam.

PUBLICATIONS

[1] **Serhan Yilmaz** and Kemal Oflazer, **"CASCADES: Compound AI Systems for Controlled And Diverse Question Generation,"** in preparation, 2025.

[2] Angelika Romanou, *[...]*, **Serhan Yilmaz**, *[...]* Sara Hooker, Antoine Bosselut, **"INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge,"** ICLR 2025 - Spotlight.

RESEARCH EXPERIENCE

Carnegie Mellon University — ***Research Internship***   *Pittsburgh, PA (Remote)* | DEC 2023 - PRESENT
- Research internship with Professor Kemal Oflazer at Carnegie Mellon University on advanced question generation and multi-agent frameworks.
- Developed a novel **Compound AI system** utilizing multiple specialized language models in structured workflows for enhanced question generation.
- Implemented sophisticated LLM-as-a-Judge methodologies with Chain-of-Thought reasoning for semantic and structural assessment.
- Designing iterative feedback loops in multi-agent collaboration frameworks, achieving over **60% improvement** over baseline generation methods.
- Authored the **CASCADES framework [1]** paper, introducing compound AI systems for controlled question generation through structured LLM orchestration and iterative refinement (in preparation, 2024).

Expedition Aya - Cohere for AI — ***Research Contributor***   *Remote* | AUG 2024 - OCT 2024
- Co-authored **INCLUDE [2]**, a multilingual benchmarking dataset for LLMs (ICLR 2025 submission).
- Built Python pipelines using LLM and VLMs for automated extraction of multichoice questions from educational resources across diverse languages.
- Involved in dataset curation with 1000+ questions and mentoring international contributors to increase diversity.

TU Darmstadt — ***Research Intern***   *Hessen, Germany (Remote)* | JUL 2024 - SEP 2024
- Developed a self-sustaining **agentic chatbot system** for psychological assessment of Ukrainian refugees under Prof. Iryna Gurevych.
- Implemented complex decision-making pipelines for autonomous question planning, contextual understanding, and verdict generation.
- Optimized the system for deployment at Charité Berlin, integrating cultural sensitivity parameters and robust assessment protocols.

KTH Royal Institute of Technology — ***Research Intern***   *Stockholm, Sweden* | JUN 2024 - SEP 2024
- Working on improving voice activity prediction models under Prof. Gabriel Skantze to create **conversational systems** capable of human-like turn-taking dialogue.
- Implemented key-value caching for real-time inference and experimented with attention sinks for higher accuracy while maintaining efficiency.
- Implemented data augmentation techniques for audio processing, enhancing model robustness in real-world scenarios.
- Successfully deployed enhanced models on Furhat Robotics' humanoid robots, improving conversational turn-taking accuracy by **93%**.

EPFL — *Summer@EPFL Intern*   *Lausanne, Switzerland* | Jun 2023 - Sep 2023
- Initiated **Project Charisius**, for developing robust, privacy-preserving Federated ML environments.
- Optimized gradient aggregators, leveraging **PyTorch** and **CUDA** to achieve performance improvements of **2200%** compared to previous versions.
- Published the Charisius library for accelerated Statistics and Federated ML applications, reducing runtime by up to **96%**.
- Developed documentation website and Flask-backed benchmarking leaderboard for the project.

Boston University — *Research Assistant*   *Boston, MA (Remote)* | Jun 2022 - Sep 2022
- Conducted research on Threat Modeling and Component Design for large-scale server systems under Prof. Rabia Tugce Yazicigil.
- Developed a C++ tool providing **336** combinations to design secure servers for resource-intensive applications.

Leadership Experience

kAi Sabanci — *Founder & President*   *Istanbul, Turkey* | Mar 2023 - Present
- Founded and led Sabanci University's AI club, growing membership to **600+ with 80% active participation**.
- Organized **17 events** in one term, with workshops on Gen AI, NLP, ML, and DL, serving as a **voluntary TA**.
- Secured kAi Sabanci's position as a **founding member** of the **NVIDIA Student Network**.
- Facilitated high-profile visits, including hosting **NVIDIA AI chief Simon See** and organizing a visit to the Turkish Presidency's Digital Transformation Office.
- Supervised **6 major projects**, including ChatSU, and NeRF campus modeling.
- Selected for the prestigious **NVIDIA Student Spotlight** series for outstanding leadership and achievement.

Work Experience

Yapi Kredi Technologies — *NLP R&D Engineering Intern*   *Istanbul, Turkey* | Oct 2023 - Apr 2024
- Enhanced document interpretation algorithms using OCR-free **Donut Transformers**.
- Fine-tuned language models, reducing annotation effort by approximately **40%** through transfer learning.

Sabanci University — *Undergraduate Teaching Assistant*   *Istanbul, Turkey* | Sep 2022 - Jan 2023
- Wrote homework assignments for students and held weekly recitation and office hours.

Pharus Tech — *Intern*   *Stockholm, Sweden (Remote)* | Jul 2020 - Sep 2020
- Analyzed and visualized data from the ESCO dataset, with information on **3008** occupations and **13890** skills.
- Developed data processing pipelines, improving data analysis efficiency by an estimated **30%**.

Awards

Sabanci University — *Sakip Sabanci Award for Outstanding Success*   Sep 2024 -
- Demonstrated academic excellence throughout the year. Awarded 50% tuition scholarship.

Sabanci University — *Sakip Sabanci Award for Outstanding Success*   Sep 2022 -
- Ranked **top 4%** in class, demonstrating academic excellence. Awarded 100% tuition scholarship.

The Royal Swedish Academy of Engineering Sciences — *Innovation in Crisis Award*   May 2020 -
- Secured **1st place** as part of team United in Crisis for innovative solutions during challenging times.

The New York Academy of Sciences — *Junior Academy Membership*   Sep 2019 -
- Selected for the prestigious "The Junior Academy" program with a highly competitive **8%** acceptance rate.

Skills

- **Languages & Tools: Python**, **Docker**, **Linux**, **AWS**, **Git**, **Javascript**, **HTML/CSS**, **C**, **C++**
- **Libraries & Frameworks: PyTorch**, **OpenCV**, **CUDA**, **Pandas**, **NumPy**, **Transformers**, **LangChain**
- **Specialized Skills:** Compound AI Systems, Multi-Agent Systems, LLM Fine-tuning, Prompt Engineering