Mohammad Reza Samsami

DEVELOPING AGENTS CAPABLE OF PLANNING IN COMPLEX SITUATIONS.

Google / Mila - Quebec AI Institute

mohammad-reza.samsami@mila.quebec | www.mrsamsami.github.io | mrsamsami | mohammadrezasamsami

Selected Research Experience

Google *Montreal, Canada*
STUDENT RESEARCHER *2024 – Pres.*
Working on world models in realistic settings in order to employ in robotics.

ServiceNow *Montreal, Canada*
VISITING RESEARCHER *2023 – 2024*
Led a project to **systematically evaluate the reasoning capabilities of LLMs through their resilience to deception** (Pub #1).
Previously focused on correcting confounding biases in the RLAIF regime to **ensure reliable post-training**.

Mila – Quebec AI Institute *Montreal, Canada*
RESEARCH STUDENT *2021 – Pres.*
Developed an RL method achieving **superhuman performance in memory-intensive tasks** (Pub #2, **Top 1.2%** @ ICLR).
Conducted **reverse engineering** of OpenAI's VPT agent to decode its behavior (Pub #3).

École Polytechnique Fédérale de Lausanne (EPFL) *Lausanne, Switzerland*
RESEARCH INTERN *2020 – 2021*
Led a project on **causal imitation learning for autonomous driving** (Pub #5).

Selected Publications

M. R. Samsami, et al. 2024. Too Big to Fool: Resisting Deception in Language Models. Under Review.

M. R. Samsami, A. Zholus, J. Rajendran, S. Chandar. 2024. Mastering Memory Tasks with World Models. ICLR, **oral**.

K. Jucys, et al. 2024. Exploratory Analysis of VPT, a Minecraft Agent. Mechanistic Interpretability at ICML.

S. Joseph, A. Zholus, **M. R. Samsami**. 2023. Mechanistic Interpretability on the Video PreTraining Agent. ATTRIB at NeurIPS.

M. R. Samsami, M. Bahari, S. Salehkaleybar, A. Alahi. 2023. Causal Imitative Model for Autonomous Driving. arXiv.

F. Hosseini, H. Fooladi, **M. R. Samsami**. 2019. Recognizing Arrow of Time in Short Stories. WiNLP at ACL.

Education

Université de Montréal *Montréal, Canada*
PH.D. IN COMPUTER SCIENCE *2022 - Pres.*

- Supervisors: Prof. Sarath Chandar, Prof. Irina Rish
- **GPA: 4.15 / 4.3**

Sharif University of Technology *Tehran, Iran*
B.SC. IN COMPUTER SCIENCE *2016 - 2021*

- **GPA: 18.18 / 20**

Selected Honors & Grants

2022, 2023	**Merit-based Scholarships for Excellence in Research**, Université de Montréal
2015	**Silver Medal in Iranian National Olympiad in Informatics**, National Elites Foundation
2016	**Ranked Top 0.3%**, Iran's Universities Entrance Exam