Google NYC
111 8th Ave
New York, NY 10011

+1(857)303-1203
aghandeharioun@google.com
alum.mit.edu/www/asma_gh
@ghandeharioun

Asma Ghandeharioun

Interests

Interpreting (Language) Models, Aligning AI with Human Values:
Why models make certain predictions, *where* they store knowledge, and *how* to constrain their generations.

Experience

Google DeepMind, Senior Research Scientist, May 2024 - Present.

Google Research, Research Scientist, Sep. 2021 - Apr. 2024.
Research Intern, Sep. 2019 - Jan. 2020.
Software Engineering Intern, Jun. 2018 - Aug. 2018.

MIT Media Lab, Research Assistant, Sep. 2014 - Jun. 2021.

Microsoft Research, Research Intern, Jun. 2017 - Aug. 2017.

Education

Ph.D. in Media Arts and Sciences, Media Lab, **MIT** (2016 - 2021); *GPA: 5.0/5.0*
Thesis: Towards Human-Centered Optimality Criteria.

M.Sc. in Media Arts and Sciences, Media Lab, **MIT** (2014 - 2016); *GPA: 5.0/5.0*

B.Sc. in Computer Engineering, **Sharif University of Tech.** (2009 - 2014)

Selected publications

See a more complete publication list on google scholar. * equal contribution. ⋄ equal advising.

Preprints

Preprints

1. Yehudai, G., Kaplan, H., **Ghandeharioun, A.**, Geva, M., Globerson, A. (2025). When Can Transformers Count to n? **Preprint**.
2. Bhalla, U., Srinivas, S., **Ghandeharioun, A.**, Lakkaraju, H. (2025). Towards unifying interpretability and control: Evaluation via intervention. **Preprint**.

Conference Papers

1. Lepori, M.A., Mozer, M., **Ghandeharioun, A.** (2025). Racing Thoughts: Explaining Large Language Model Contextualization Errors. **NAACL**.
2. **Ghandeharioun, A.***, Yuan, A.*, Guerard, M., Reif, E., Lepori, M. A., Dixon, L. (2024). Who's asking? User personas and the mechanics of latent misalignment. **NeurIPS**.
3. **Ghandeharioun, A.***, Caciularu, A.*, Pearce, A., Dixon, L., Geva, M. (2024). Patchscopes: A unifying framework for inspecting hidden representations of language models. **ICML**.
4. Friedman, D., Lampinen, A. K., Dixon, L., Chen, D., **Ghandeharioun, A.** (2024). Interpretability illusions in the generalization of simplified models. **ICML**.
5. Hase, P., Bansal, M., Kim, B., **Ghandeharioun, A.** (2023). Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models, **NeurIPS (Spotlight - top 3%)**.
6. Krishna, S., Ma, J., Slack, D., **Ghandeharioun, A.**, Singh, S., Lakkaraju, H. (2023). Post hoc explanations of language models can improve language models. **NeurIPS**.
7. **Ghandeharioun, A.**, Kim, B., Li, C., Jou, B., Eoff, B. , Picard, R. (2022). DISSECT: Disentangled Simultaneous Explanations via Concept Traversals, **ICLR**.
8. Jaques, N.*, Shen, J.*, **Ghandeharioun, A.**, Ferguson, C., Lapedriza, A., Jones, N., Gu, S., Picard, R. (2020). Human-centric dialog training via offline reinforcement learning. **EMNLP (Oral)**.
9. Saleh A.*, Jaques N.*, **Ghandeharioun, A.**, Shen, J., Picard, R. (2020). Hierarchical Reinforcement Learning for Open-Domain Dialog, **AAAI (Oral)**.
10. **Ghandeharioun, A.***, Shen, J.*, Jaques N.*, Ferguson, C., Jones, N., Lapedriza, A., Picard, R. (2019). Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems. **NeurIPS**.

11. **Ghandeharioun, A.**, McDuff, D., Czerwinski, M., Rowan, K. (2019). EMMA: An Emotion-Aware Wellbeing Chatbot. **ACII, IEEE**.

Workshop Papers

1. Hussein, N.*, **Ghandeharioun, A.***, Hussein, N., Mullins, R., Reif, E., Wilson, J., Thain, N.$^{\diamond}$, Dixon, L$^{\diamond}$. (2024). Can Large Language Models explain Their Internal Mechanisms? **IEEE VISxAI**
2. Pearce, A.*, **Ghandeharioun, A.***, Hussein, N., Thain, N., Wattenberg, M., Dixon, L. (2023). Do machine learning models memorize or generalize. **IEEE VISxAI (Best paper).**
3. Friedman, D., Lampinen, A. K., Dixon, L., Chen, D., **Ghandeharioun, A.** (2023). Comparing Representational and Functional Similarity in Small Transformer Language Models. *UniReps: the First Workshop on Unifying Representations in Neural Models*, **NeurIPS Workshop (Oral)**.
4. Jaques N., **Ghandeharioun, A.**, Shen, J., Ferguson, C., Jones, N., Lapedriza, A., Gu, S., Picard, R. (2019). Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog, *Conversational AI*, **NeurIPS workshop**.

Professional Service

Chair/Organizer: R2HCAI: Representation Learning for Responsible Human-Centric AI, AAAI, 2023.

Senior Area Chair, 2024-present:

- **NeurIPS**.

Area Chair/Senior Program Committee, 2023-present:

- **NeurIPS (Outstanding AC)**, **ICLR**, **COLM**, **ACII**.

Reviewer/Program Committee, 2015-present:

- **Machine Learning**: **NeurIPS**, **ICML**, **ICLR**, **AAAI**, **IJCAI**.
- **ML Applications**: ACII, JBHI.
- **Human-Computer Interaction**: CHI **(Excellent Reviewer)**, DIS, TOCHI, Psychology of Well-Being Journal, IMWUT, UbiComp.

Mentorship

Mentor for the Google PhD Fellowship recipients, Ziming Liu, Shan Chen, 2024.

Mentor, interpretability & explainability round table, Women in machine learning workshop, ICML 2024.

Mentor for PhD interns Peter Hase, Dan Friedman, Michael Lepori, 2021-present.

gMentor on internal Google mentorship program, 2022-present.

Advisor MIT Alumni Advisors Hub, 2021-present.

Mentor for Master of Engineering (MEng) and Undergraduate Research Opportunities Program (UROP): Darian Bhathena, Alexander Lynch, Diane Zhou, Marek Subernat, 2016-2021.

Students Offering Support: Assisting underrepresented students applying to the Media Lab, 2017.

Honors and Awards

2.5M grant from **NIH**, 5R01MH118274, 2019-2023 (PIs: P. Pedrelli, R. Picard).

150K grant from **J-Clinic** for conducting machine learning in healthcare research, 2019 (PI: R. Picard).

D. E. Shaw Zenith Fellowship, 2021.

MIT Quest for Intelligence, MIT Stephen A. Schwarzman College of Computing, Machine Learning Across Disciplines Challenge, recipient of unlimited Google Cloud Platform credit, 2019.

Silver Medal in Iranian National Olympiad in Informatics, 2008.

National Elites Foundation Grant recipient for outstanding academic success in undergraduate studies in Iran, 2009 - 2014.