Eghbal A. Hosseini

ehoseini@mit.edu https://eghbalhosseini.github.io
eghbalhosseini Eghbal A. Hosseini

My research sits at the intersection of computational neuroscience and AI, studying how data and context shape internal representations and behavior in Large Language Models. Prior work in mechanistic interpretability covers representational geometry during pre-training and in-context learning, and convergence across models, modalities, and the brain. Current focus: pre- and post-training strategies for agentic behavior and long-term planning in LLMs.

Current Position

2025 – Present **Visiting Researcher**, Google DeepMind, San Francisco, CA
Research on how context and in-context learning reshape the representational geometry of LLMs.

Education

2016 – 2024 **Ph.D. Computational Neuroscience**, Massachusetts Institute of Technology (MIT)
Thesis: *Towards Synergistic Understanding of Language Processing in Biological and Artificial Systems*.

2023 **Analytical Connectionism**, Gatsby Computational Neuroscience Unit, University College London.

2017 **Brains, Minds, and Machines**, Marine Biological Laboratory University of Chicago.

2012 – 2014 **M.Sc. Electrical Engineering**, George Mason University (GMU)

2005 – 2010 **B.Sc. Electrical Engineering**, Iran University of Science and Technology (IUST)

Research Publications

Selected

King, J., Fedorenko, E., & **Hosseini**†, **E. A.** (2026). Representational curvature modulates behavioral uncertainty in large language models [†senior author]. *International Conference on Machine Learning (ICML)*.

Hosseini, **E. A.**, Cheung, B., Fedorenko, E., & Williams, A. H. (2026). Modulating cross-modal convergence with single-stimulus, intra-modal dispersion. *ICLR 2026 Re-Align Workshop*.

Hosseini, **E. A.**, Li, Y., Bahri, Y., Campbell, D., & Lampinen, A. K. (2026). Context structure reshapes the representational geometry of language models. *arXiv*.

Lampinen, A. K., Li, Y., **Hosseini**, **E. A.**, Bhardwaj, S., & Shanahan, M. (2026). Linear representations in language models can change dramatically over a conversation. *arXiv*.

Hosseini, **E. A.**, Casto, C., Zaslavsky, N., Conwell, C., Richardson, M., & Fedorenko, E. (2024). Universality of representation in biological and artificial neural networks. *bioRxiv*.

Hosseini, **E. A.**, Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiol Lang (Camb)*, *5*(1), 43–63.

Hosseini, **E. A.**, & Fedorenko, E. (2023). Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *NeurIPS*.

Schrimpf, M., Blank*, I. A., Tuckute*, G., Kauf*, C., **Hosseini**, **E. A.**, Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing [*equal contribution]. *Proc. Natl. Acad. Sci. U. S. A.*, *118*(45).

Additional

Cohen*, Z., Jagadish*, A. K., **Hosseini***, **E. A.**, & Eckstein, M. K. (2026). Reinforcement learning: Computational modeling of learning and decision-making [*equal contribution]. *Proceedings of the Analytical Connectionism Schools 2023–2024, PMLR, 320*, 87–101. 🔗.

Regev, T. I., Casto, C., **Hosseini**, **E. A.**, Adamek, M., Ritaccio, A. L., Willie, J. T., Brunner, P., & Fedorenko, E. (2024). Neural populations in the language network differ in the size of their temporal receptive windows. *Nat. Hum. Behav.*, *8*(10), 1924–1942. 🔗.

Hosseini, **E. A.**, Zaslavsky, N., Casto, C., & Fedorenko, E. (2023). Teasing apart the representational spaces of ANN language models to discover key axes of model-to-brain alignment. *Computational Cognitive Neuroscience, (Oral presentation, top 5% submission).* 🔗.

Wang, J., **Hosseini**, **E. A.**, Meirhaeghe, N., Akkad, A., & Jazayeri, M. (2020). Reinforcement regulates timing variability in thalamus. *Elife*, *9*. 🔗.

Remington, E. D., Narain, D., **Hosseini**, **E. A.**, & Jazayeri, M. (2018). Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron*, *98*(5), 1005–1019.e5. 🔗.

Wang*, J., Narain*, D., **Hosseini**, **E. A.**, & Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses [*equal contribution]. *Nat. Neurosci.*, *21*(1), 102–110. 🔗.

Hosseini, **E. A.**, Nguyen, K. P., & Joiner, W. M. (2017). The decay of motor adaptation to novel movement dynamics reveals an asymmetry in the stability of motion state-dependent learning. *PLoS Comput. Biol.*, *13*(5), e1005492. 🔗.

Talks & Presentations

Invited Talks

Towards Synergistic Understanding of Language Processing in Biological and Artificial Systems

Feb 2026	Sungkyunkwan University (SKKU), Suwon, South Korea
Apr 2025	Department of Psychology, UC Berkeley
Dec 2024	ANCOR Seminar, Computer Science, Brown University
Sep 2024	Google DeepMind, Mountain View, CA

Conference & Workshop Talks

Aug 2025	**Universality of representation across biological and artificial neural networks**, Community Event: "Universality and Idiosyncrasy of Perceptual Representations", Cognitive Computational Neuroscience (CCN) 2025, University of Amsterdam, Netherlands
Aug 2023	**Teasing apart the representational spaces of ANN language models to discover key axes of model-to-brain alignment** (oral presentation, top 5% submission), Cognitive Computational Neuroscience (CCN) 2023, Oxford, UK

Technical & Research Skills

Research Expertise	Language modeling, Representation learning, Mechanistic interpretability, Computational neuroscience, Neuroimaging analysis (fMRI, ECoG)
Programming & ML Frameworks	Python, JAX/Flax, PyTorch, TensorFlow, Slurm, MATLAB, R, LaTeX
Languages	Fluent in reading, writing, and speaking *English* and *Persian*

Experiences & Awards

Research Experience

2025–Present **Visiting Researcher**, Google DeepMind, San Francisco, CA

Spring 2025 **Visiting Scientist**, Special Year on Large Language Models and Transformers, Part 2, Simons Institute for the Theory of Computing, UC Berkeley

2024–2025 **Postdoctoral Associate**, Dr. Evelina Fedorenko, EvLab, McGovern Institute for Brain Research, MIT

2019–2024 **Graduate Research Assistant**, Dr. Evelina Fedorenko, EvLab, McGovern Institute for Brain Research, MIT

2017–2018 **Graduate Fellowship Student**, Dr. Edward S. Boyden, Synthetic Neurobiology Group, McGovern Institute for Brain Research, MIT

2015–2016 **Technical Assistant**, Dr. Mehrdad Jazayeri, JazLab, McGovern Institute for Brain Research, MIT

2013–2014 **Graduate Research Assistant**, Dr. Wilsaan Joiner, Sensorimotor Integration Lab, Volgenau School of Engineering, GMU

Mentoring

2024–2026 **Jack G. King**, MIT SuperUROP — mentored an undergraduate as senior author on a first-authored ICML 2026 paper on representational curvature and behavioral uncertainty in LLMs

2020–2022 **Alexandra So**, MIT UROP — optimal stimulus design for neural network models of language processing

Teaching Experience

2017–2020 **Teaching Assistant**, MIT — Introduction to Neural Computation (M. Fee); Science of Intelligence (T. Poggio)

Reviewing

2024–Present NeurIPS, ICLR, ICML (conferences); Nature Human Behaviour (journal)

Honors & Awards

2023 **Oral presentation (top 5% of submissions)**, Cognitive Computational Neuroscience (CCN)

2020 **Friends of the McGovern Institute Fellowship**, MIT

2017–2018 **BCS Hilibrand Graduate Student Fellowship**, MIT

2016–2017 **Henry E. Singleton(1940) Presidential Fellowship**, MIT