

Kerem Oktar, Ph.D

609-375-4325 | oktar.research@gmail.com | keremoktar.com

SUMMARY

AI researcher at Meta FAIR developing evaluations and post-training methods focused on safety and alignment. Princeton Ph.D. in cognitive science; I use computational models and experiments to improve model behavior.

EXPERIENCE

Meta Fundamental AI Research (FAIR)

2025 – Present

Postdoctoral Researcher

Seattle, WA

- Co-built an end-to-end evaluation pipeline for **sycophancy in frontier models**: designed a synthetic multi-turn conversation generator, developed an LLM-as-judge scoring framework, and calibrated it against both lay and expert judgment (appears in Meta’s frontier model [safety report](#)).
- Created a benchmark of **LLMs’ ability to infer users’ beliefs and preferences** from conversation; managed vendor engineers and data scientists to collect >10,000 conversational turns; surfaced a >20% spread across frontier models.
- Built synthetic data pipelines for SFT and RL on **multi-agent orchestration** tasks; partnered with post-training teams on data-quality standards, prompt design, and downstream alignment.
- Monitored and refined evaluation tools: contributed a **reasoning-robustness** analysis to [BrittleBench](#), showing small prompt perturbations degrade performance up to 12% and reverse model rankings.
- Currently developing an **automated alignment research system** for improved model honesty and safety.

Princeton University

Princeton, NJ

Doctoral Researcher, Computational Cognitive Science Lab

2019 – 2025

- Led end-to-end research programs on **how people learn, persuade, and change their minds across >100 experiments and >10 papers**: translated open-ended questions into measurable designs, built interactive UIs, and analyzed behavioral data at scale with rigorous statistical methods.
- Developed **Bayesian computational models** of how agents should weigh communicator motives (self-serving vs. benevolent) and applied them to LLM evaluation, improving predictive accuracy by 40%.
- Delivered **>20 invited and conference talks** (Harvard, MIT, Stanford, UW CSE, Max Planck) to technical and general audiences; designed and taught courses; received tenure-track offer.

EDUCATION

Princeton University

Princeton, NJ

Ph.D. @ Computational Cognitive Science Lab (Advisors: Tania Lombrozo & Tom Griffiths)

2019-2025

Pomona College

Claremont, CA

B.A. in Economics and Cognitive Science, *cum laude, honors*

2015-2019

SELECTED PUBLICATIONS (h-index: 13; 663 citations)

Wu*, Liu*, **Oktar***, Sumers & Griffiths, 2025. Are LLMs Sensitive to Motives Behind Communication? *NeurIPS*
Oktar & Lombrozo, 2025. How Beliefs Persist Amid Controversy: The Paths to Persistence Model. *Psych. Rev.*

TECHNICAL SKILLS

AI Engineering: Python, R; prompt design; agentic architectures; post-training; AI-assisted development.

Statistics & Social Science: experimental and survey design, causal inference, mixed-effects regression, power analysis, large-scale behavioral data collection; formal modeling of mental processes.

AWARDS & SERVICE

Awards: Marr Prize for Best Paper (CogSci); Rockefeller Fellowship (\$110K); Princeton Teaching Award.

Reviewing: *NeurIPS*, *ICML*, *ICLR*, *TACL*, *Nature Machine Intelligence*, *Psychological Science*, *Cognition*.