Our models are as ethical as we are!

Models do not mirror humans,
yet they can amplify, suppress, or transform our prejudice.

THESIS STATEMENT

Social bias has long shaped human society, and decades of social psychology offer ways to identify and reduce it.

By positioning social psychology as a foundation for bias research in NLP, my dissertation advances beyond narrow benchmark-driven association tests and targeted debiasing approaches to

(1) **discover** hidden stereotypes,

(2) **evaluate** how such stereotypes manifest in model behavior,

(3) develop generalizable **mitigation** strategies, and

(4) ultimately move beyond isolated model behavior to **agentic systems** toward more responsible and trustworthy NLP.

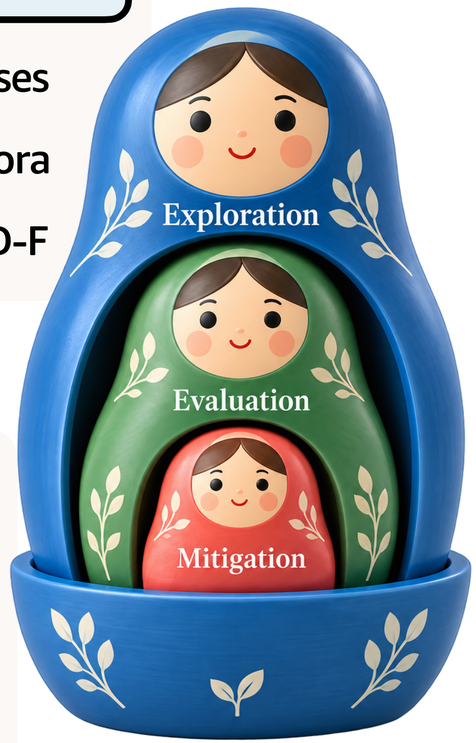
My Russian Doll Model

1. BIAS DISCOVERY

EMNLP '23 Main — *Global Voices, Local Biases

EMNLP '24 Findings — *BiasDora

AAAI '26 — BAD-F



2. BEHAVIORAL EVALUATION

ACL '26 Findings — *Talent or Luck

ACL '26 Main — *Vignette

TrustNLP '26 — Purdah and Patriarchy

EMNLP '25 Findings — What's Not Said Still Hurts

*Breaking Bias — AIES '24

*Debias It Yourself — Submission (EACL)

Towards Inclusive LMs — EMNLP '25 Findings

3. BIAS MITIGATION

*(Paper) = First authored

My Research Landscape