What differentiates in-group and out-group speech?

Counterfactual probing reveals that **affect** and **specificity** vary systematically, but in different ways, with intergroup relationship (in-group or out-group).

We found **no interaction** between the two, as hypothesized *generalizing* from the Linguistic Intergroup Bias.

Counterfactual Probing for the Influence of Affect & Specificity on Intergroup Bias

Venkata S Govindarajan, Kyle Mahowald, David I. Beaver, Junyi Jessy Li

INTERGROUP BIAS

In Govindarajan et al. (2023), we introduced the study of intergroup bias, a novel framing of bias that directly models intergroup relationships in interpersonal language. Inspired by work on the Linguistic Intergroup bias hypothesis, we investigate if 2 pragmatic features **can explain the differences** between in-group and out-group language:

Affect is a coarse grained feature that estimates how a speaker *feels* towards the target they mentioned in an interpersonal utterance.

Specificity measures the level of detail and involvement of concepts, objects and events.

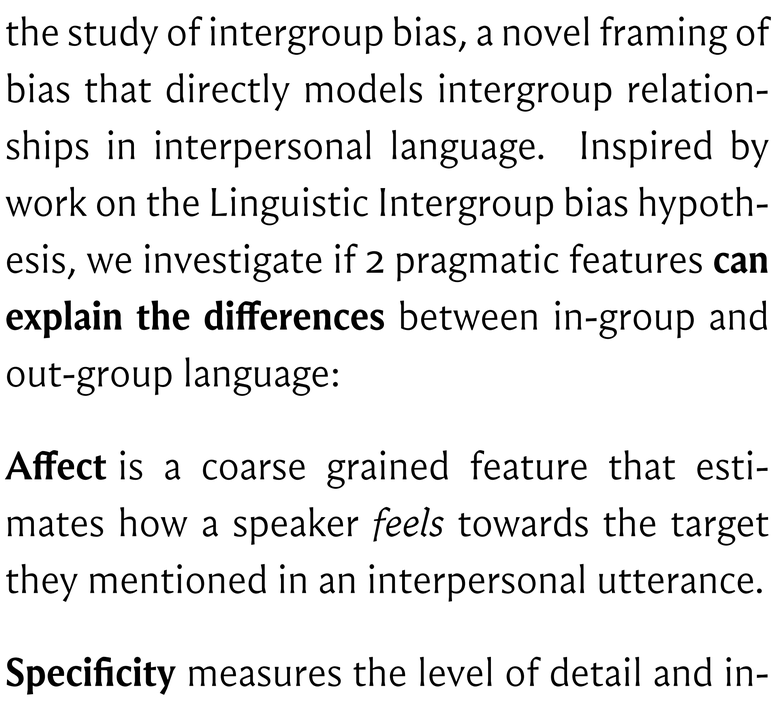

FIGURE 1: Predicted variation in language in our Intergroup bias formulation.

EXPERIMENTS

FIGURE 2: Flowchart describing the specificity intervention experiment and expected results.

To investigate if specificity and affect are **causal** explanations for intergroup bias, we probe whether a neural model finetuned for in-group vs. out-group prediction *uses* specificity or affect in its decision-making process using **AlterRep** — a counterfactual probing technique that tests if a neural network uses a property, rather than just testing if the model's learned representations correlate with the property.

Our hypothesis Interventions towards higher specificity should induce the model to predict positive affect tweets as out-group and negative affect tweets as in-group, while interventions towards lower specificity should affect the model conversely.

RESULTS

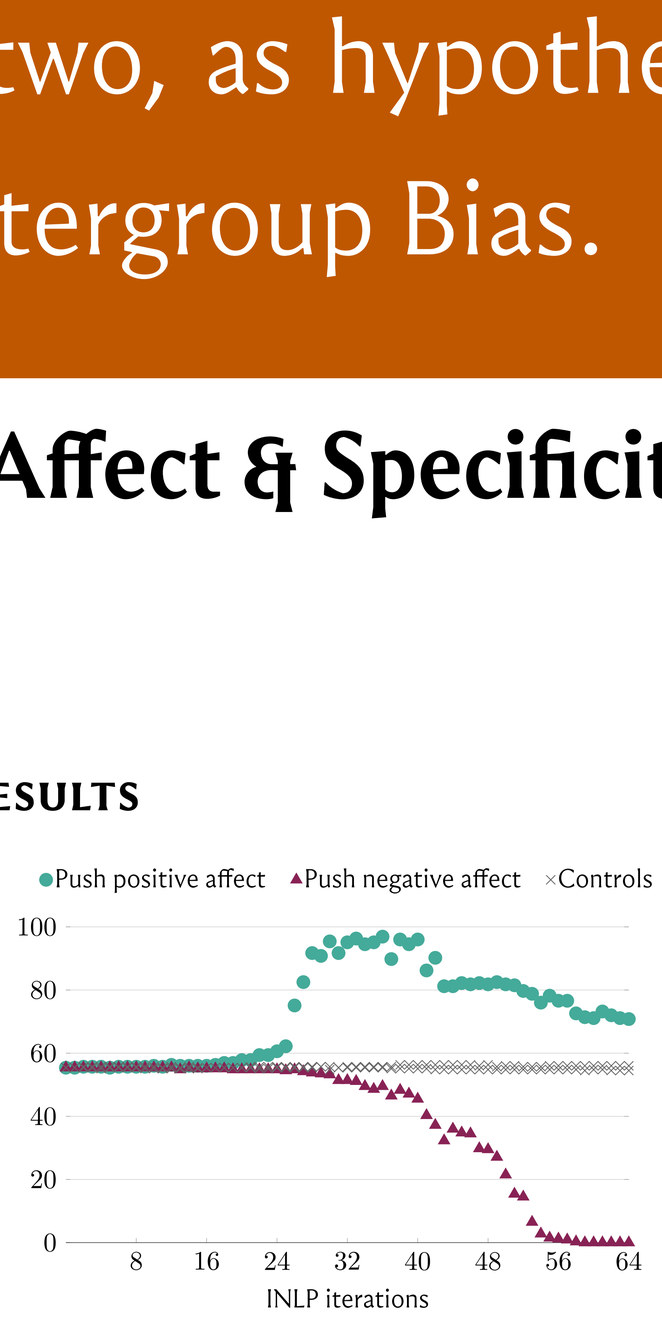

FIGURE 3: Results of affect intervention.

- We find that affect influences model predictions as we expected, but specificity interventions were causal only in the more specific direction — compare and contrast intervention effects between Figures 3 and 4.
- **No interaction** was found between specificity and affect. Intervening on one feature affected all datapoints in the test set similarly.

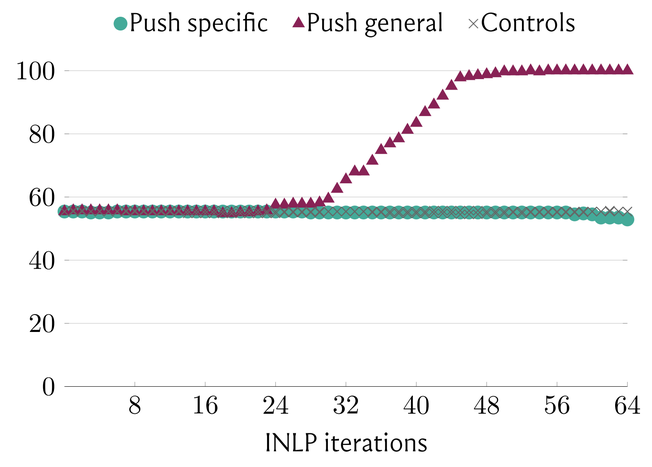

FIGURE 4: Results of specificity intervention.

Data & code available online: github.com/venkatasg/intergroup-probing

TEXAS
The University of Texas at Austin