ANNOTATION SENSITIVITY:

TRAINING DATA COLLECTION METHODS AFFECT MODEL PERFORMANCE

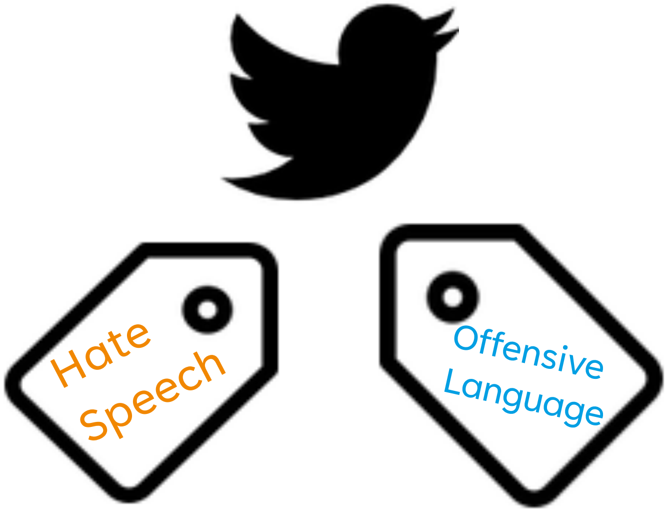
Stephanie Eckman

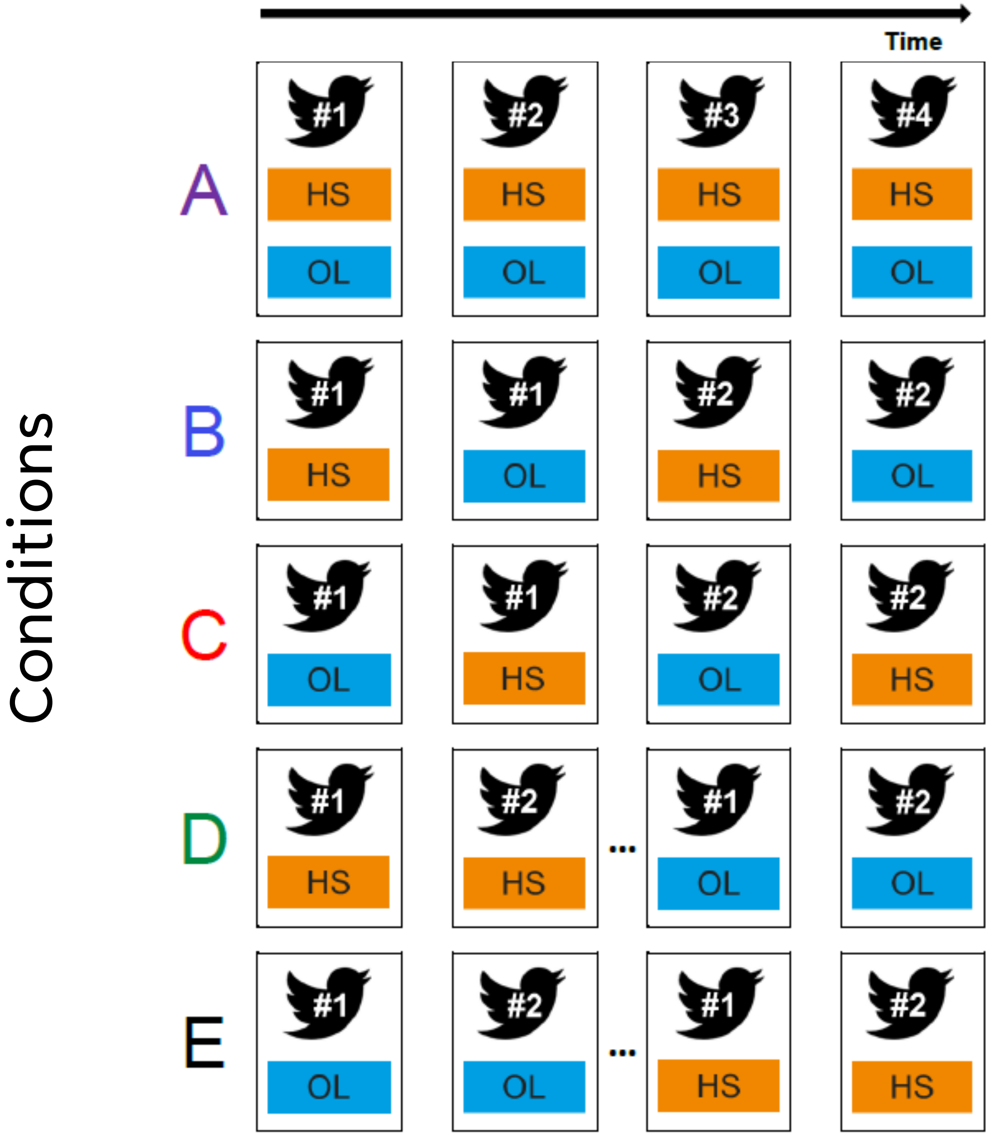
Christoph Kern, Jacob Beck, Bolei Ma, Rob Chew, Frauke Kreuter

“The bias I am most nervous about is the bias of the human feedback raters“

Sam Altman
March 25 2023 “The Lex Fridman Podcast“

RESEARCH DESIGN





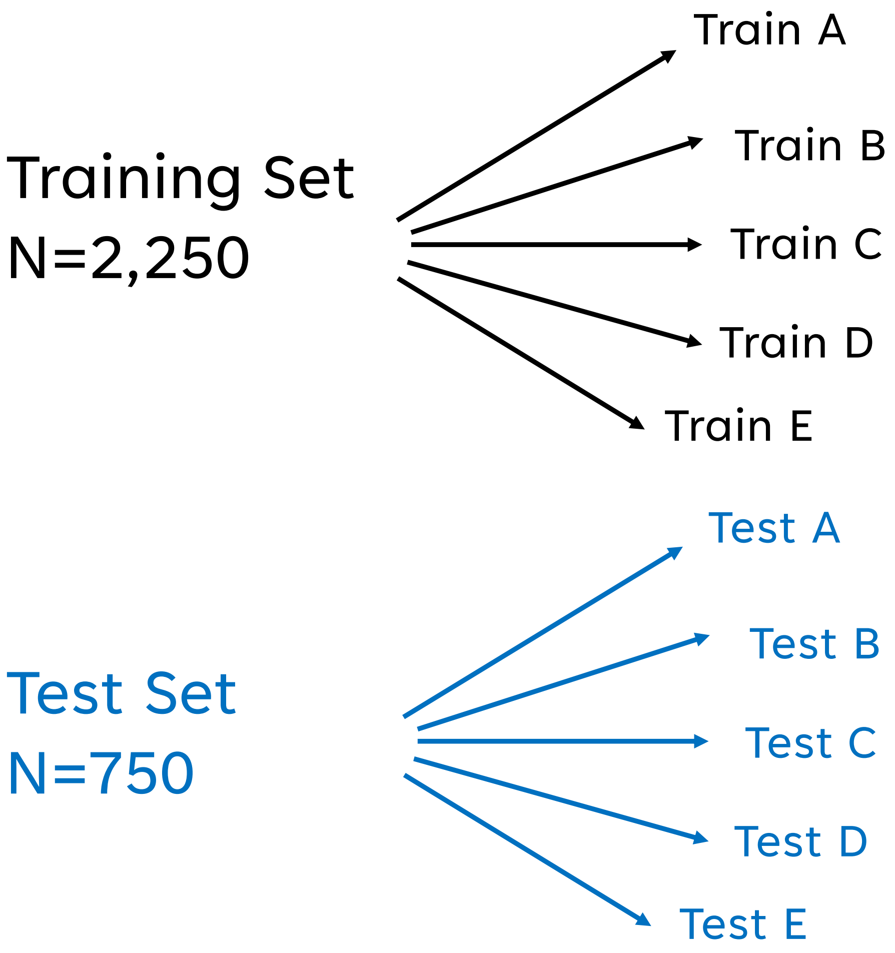
A	#1 HS OL	#2 HS OL		#3 HS OL	#4 HS OL
B	#1 HS	#1 OL		#2 HS	#2 OL
C	#1 OL	#1 HS		#2 OL	#2 HS
D	#1 HS	#2 HS	...	#1 OL	#2 OL
E	#1 OL	#2 OL	...	#1 HS	#2 HS

DATA COLLECTION

- 3000 tweets (Davidson et al 2017)
- ~900 annotators from Prolific (Nov-Dec 2022)

- 50 tweets / annotator
- 3 annotations / tweet - condition
- 15 total annotations / tweet

MODEL TRAINING



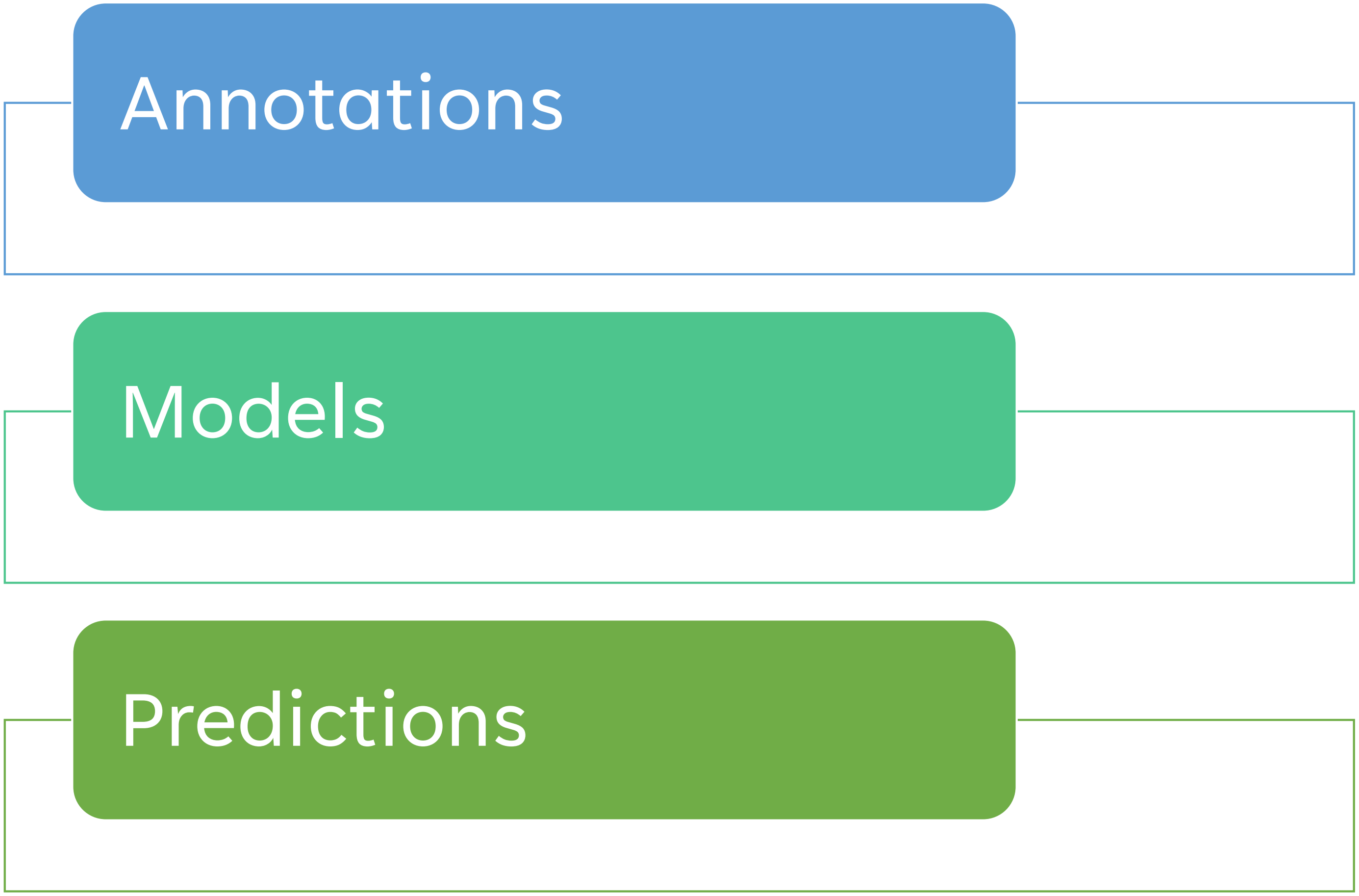
Training Set
N=2,250

Train A
Train B
Train C
Train D
Train E

Test Set
N=750

Test A
Test B
Test C
Test D
Test E

3 TYPES OF RESULTS

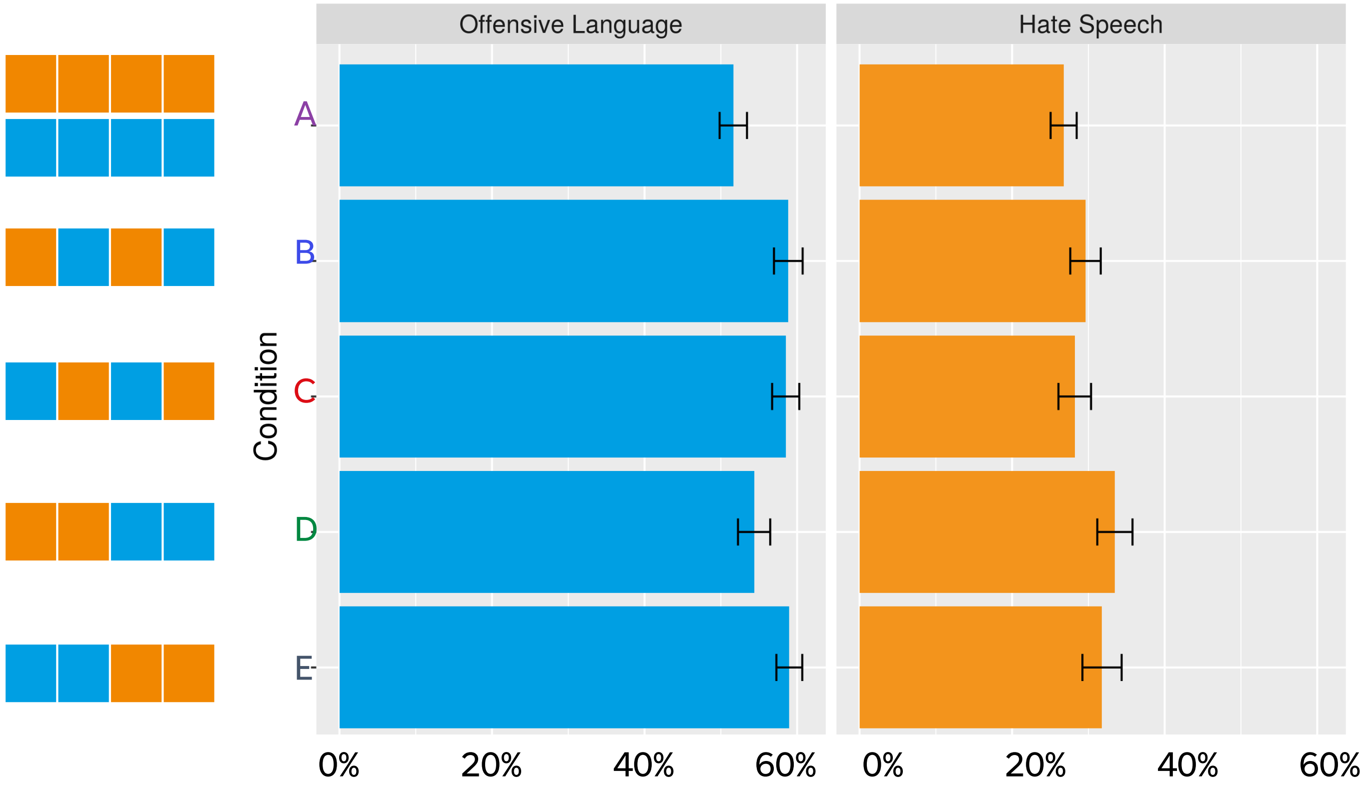


Annotations

Models

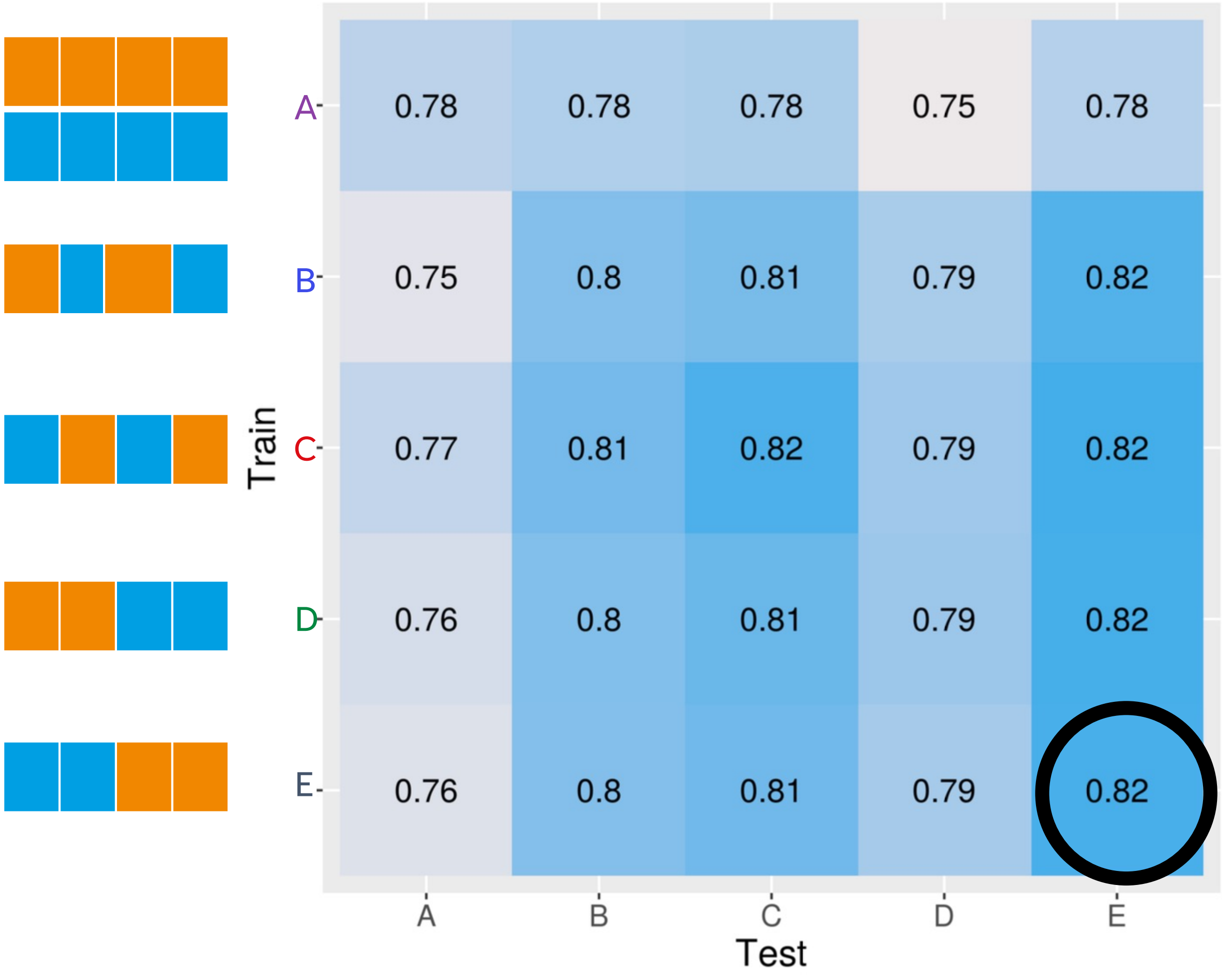
Predictions

ANNOTATIONS



MODEL PERFORMANCE

- BERT models of offensive language
- Number shown is *balanced accuracy*

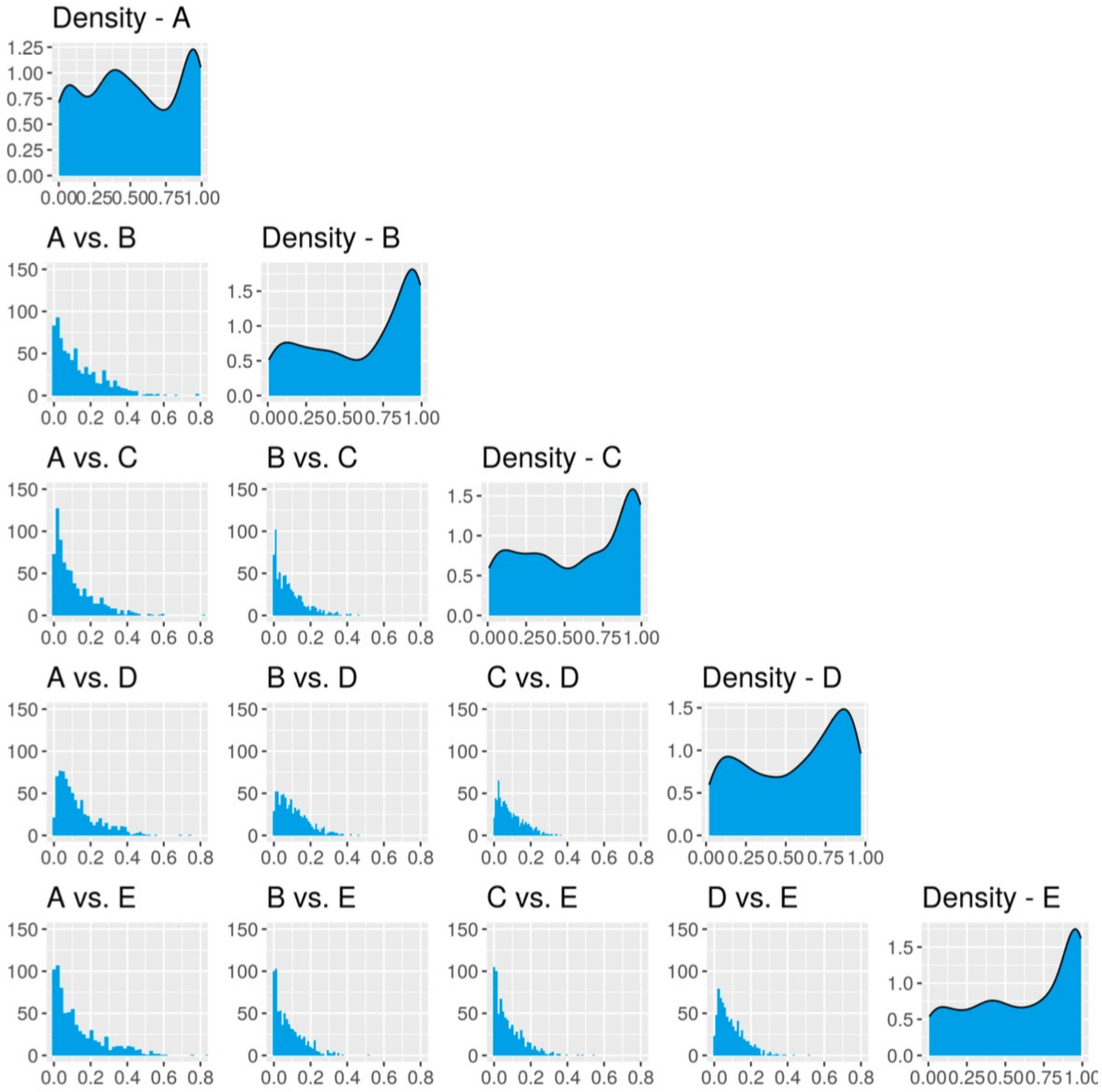


	Train	A	B	C	D	E
🟧🟧🟧🟧 🟦🟦🟦🟦	A	0.78	0.78	0.78	0.75	0.78
🟧🟦🟧🟦	B	0.75	0.8	0.81	0.79	0.82
🟦🟧🟦🟧	C	0.77	0.81	0.82	0.79	0.82
🟧🟧🟦🟦	D	0.76	0.8	0.81	0.79	0.82
🟦🟦🟧🟧	E	0.76	0.8	0.81	0.79	0.82

PREDICTIONS

- Diagonal:
  - distribution of annotations by conditions

- Off-diagonal:
  - Absolute difference of predictions



TAKEAWAYS

- How you collect annotations matters
  - for labels, models, predictions

- Some conditions perform better/worse as train/test data
  - More research needed to inform best practices

- Some evidence of fatigue
  - Fewer offensive speech labels in Condition D 🟧🟧🟦🟦
  - Fewer hate speech labels in Condition E 🟦🟦🟧🟧

THANK YOU

Stephanie Eckman

Social Data Science Center, University of Maryland

steph@umd.edu

stepheckman.com