BRINGING SURVEY METHODOLOGY TO MACHINE LEARNING

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“

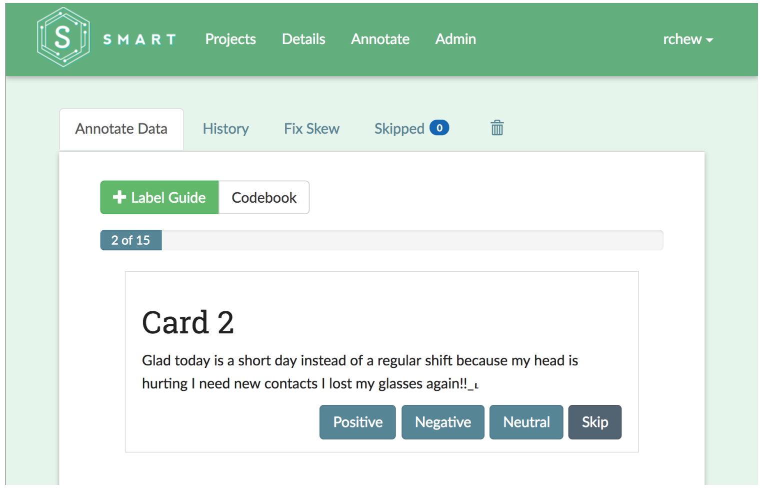
Data Collection

Error Sources

Impact

- Nonresponse
- Order Effects
- Interviewer Effects

Bias

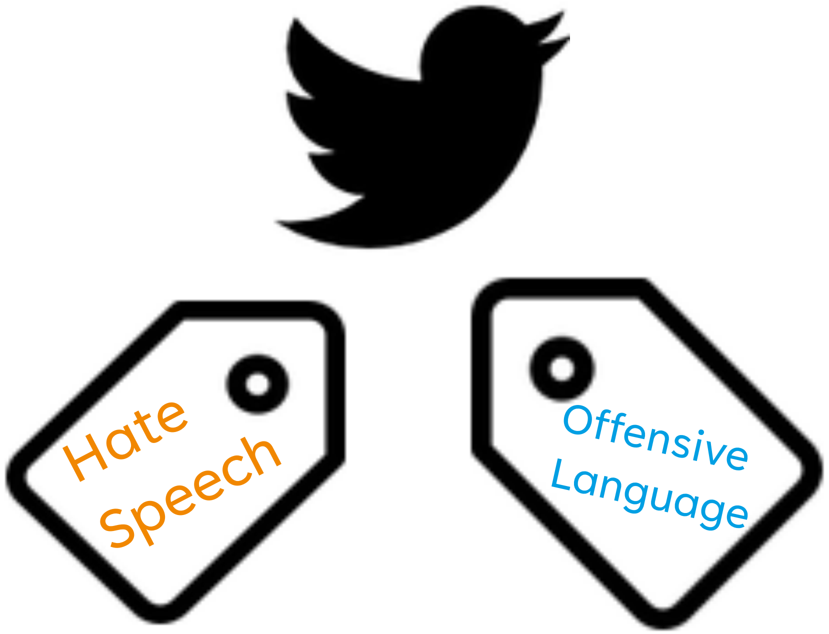


Beck et al (2022):

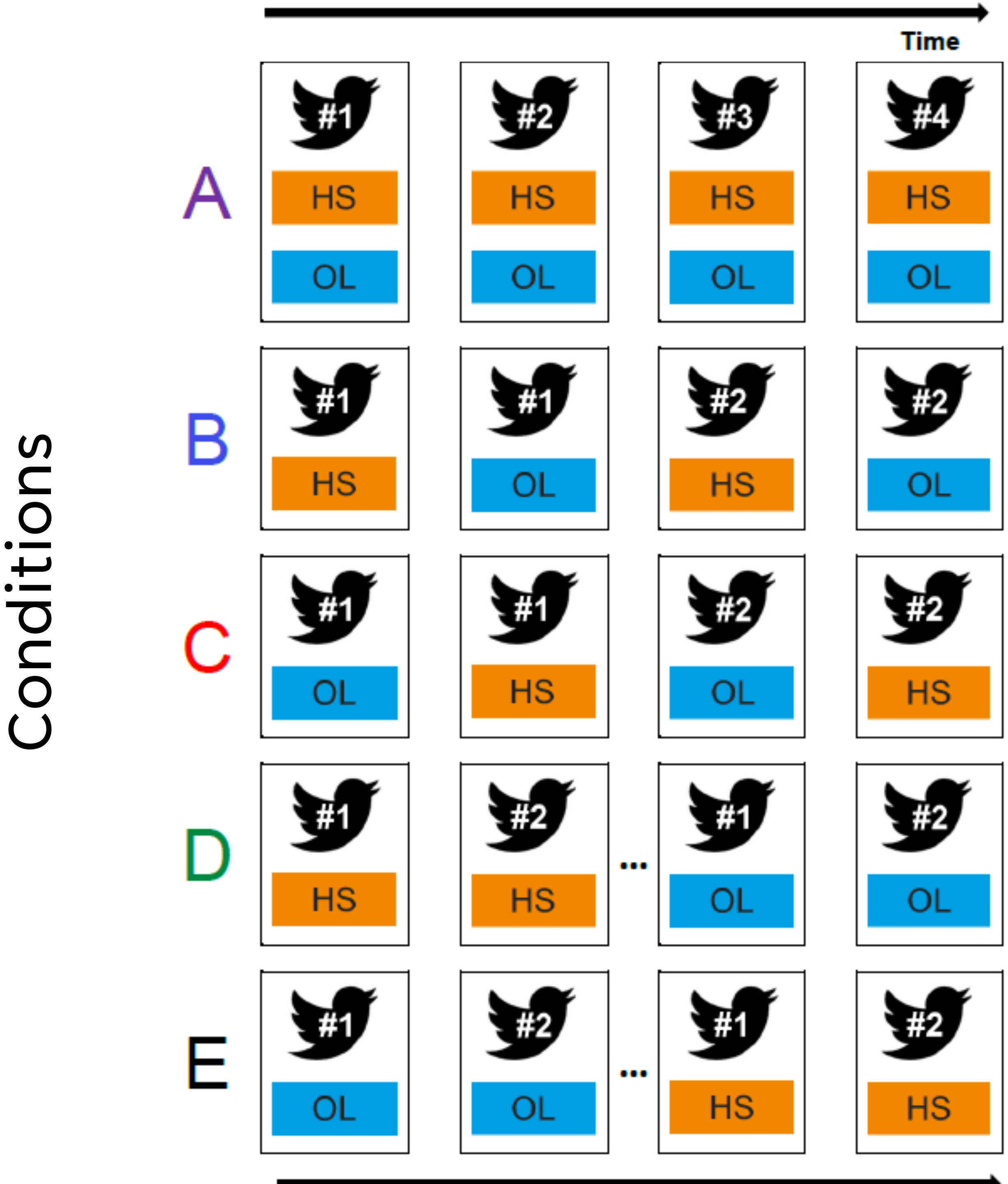
- Wording Effects
- Order Effects
- Annotator Effects

Prediction Error

RESEARCH DESIGN



- 5 annotation conditions
- Do they lead to different labels?
- Do they lead to different ***models***?


Time
Conditions
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 per tweet & condition

  →15 total annotations per tweet

→~45k annotations

MODEL TRAINING

Training Set
N=2,250

Train A
Train B
Train C
Train D
Train E

Test Set
N=750

2 model types:

- LSTM
- BERT

2 DVs:

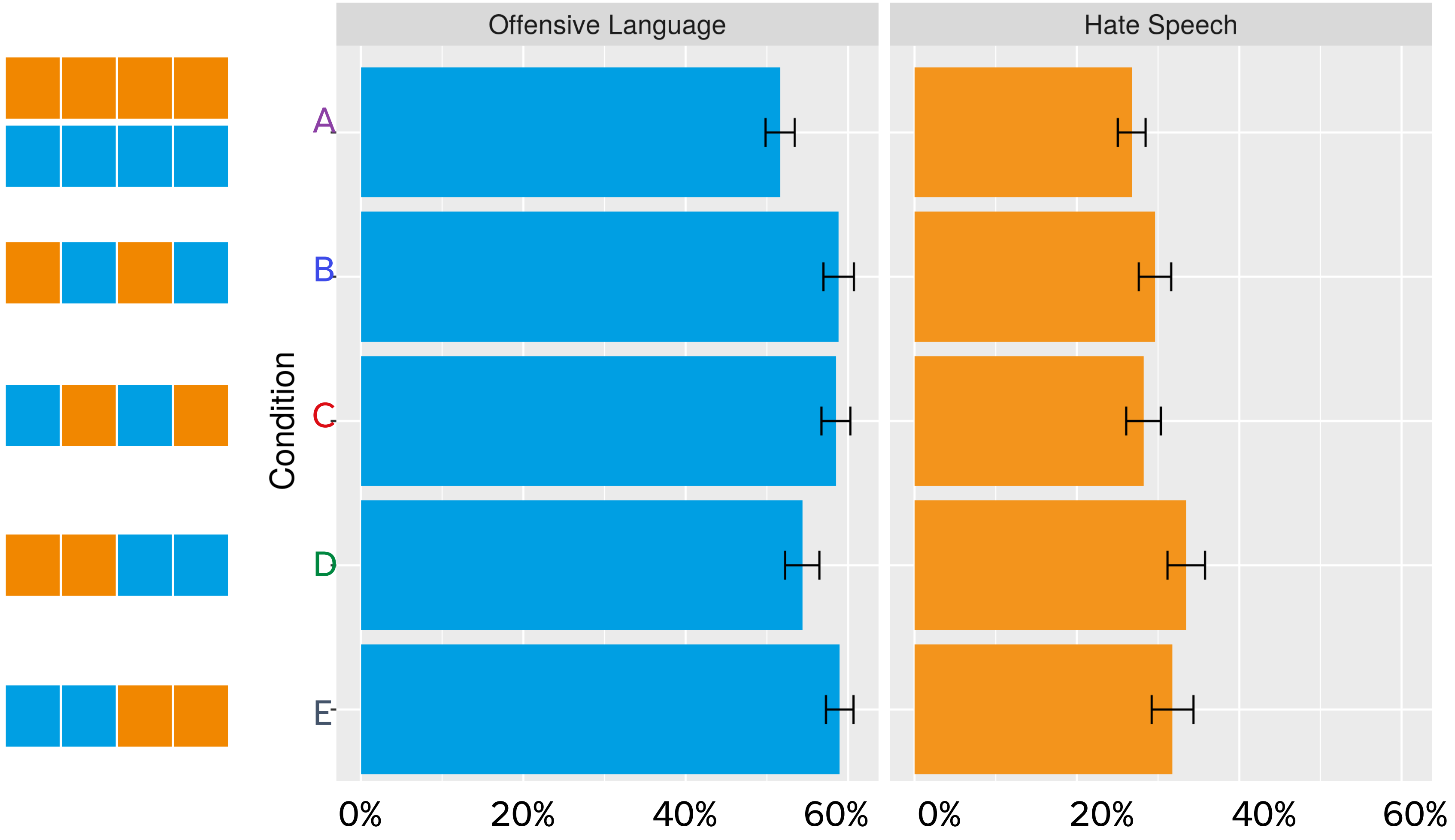
- Hate speech
- Offensive language

SETUP OF RESULTS

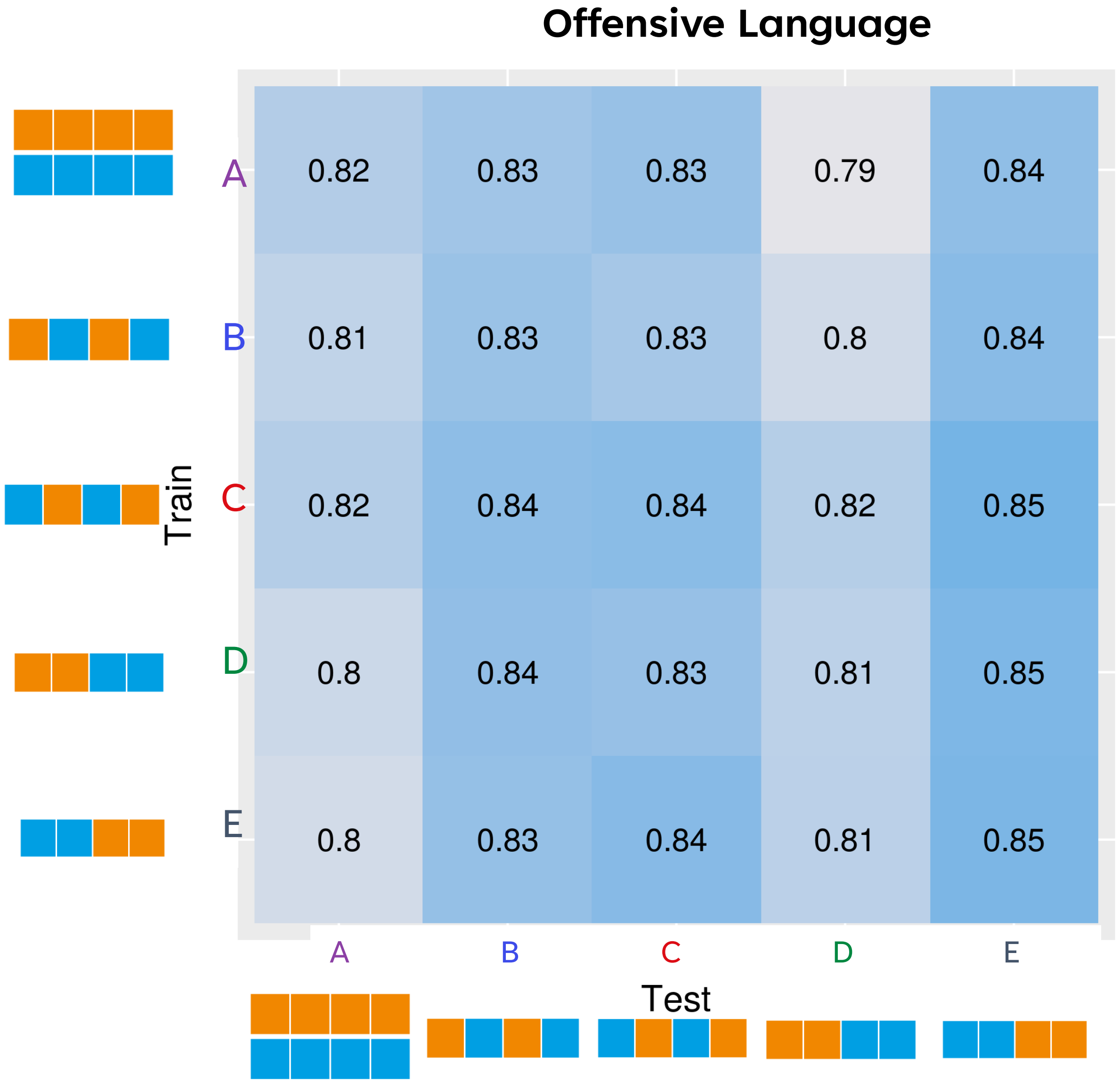
- Labels
- Model Performance
  - ROC AUC
  - Learning Curves
- Predictions

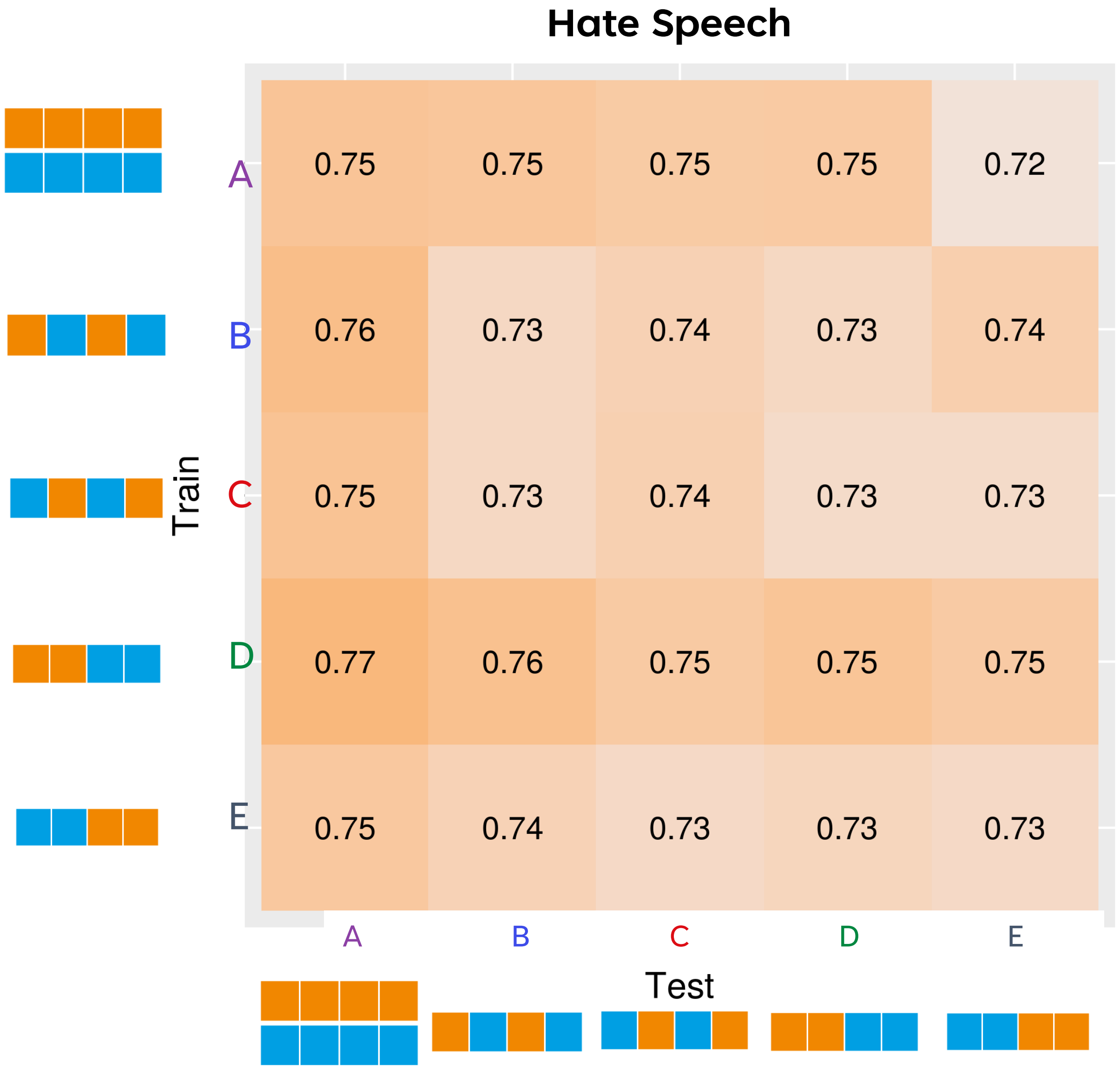
- By Condition (5)
- By Dependent Var (2)
  - Hate speech
  - Offensive language
- By model type (2)
  - LSTM
  - BERT

% OFFENSIVE LANGUAGE / HATE SPEECH BY CONDITION



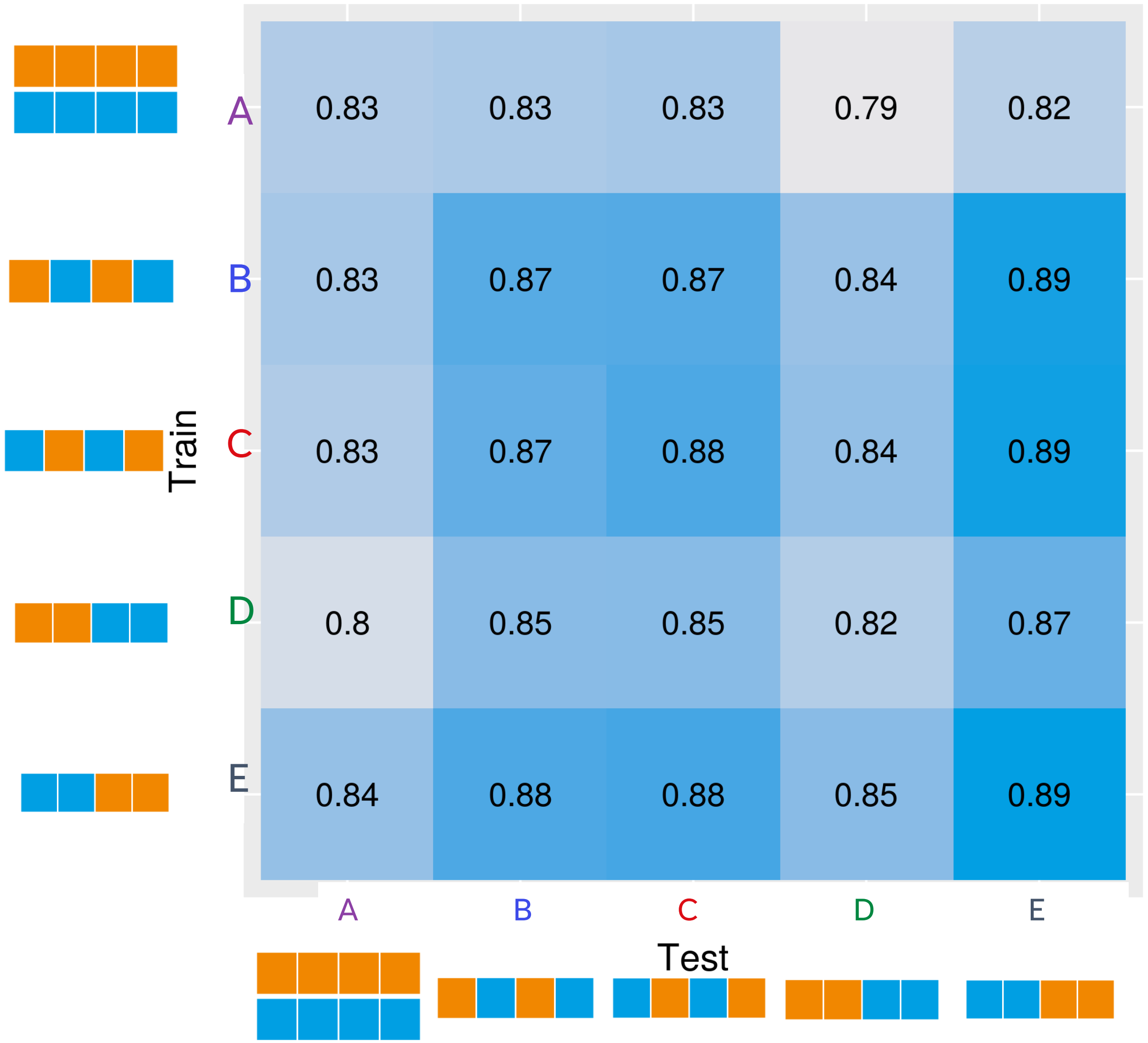
MODEL PERFORMANCE ROC-AUC: LSTM



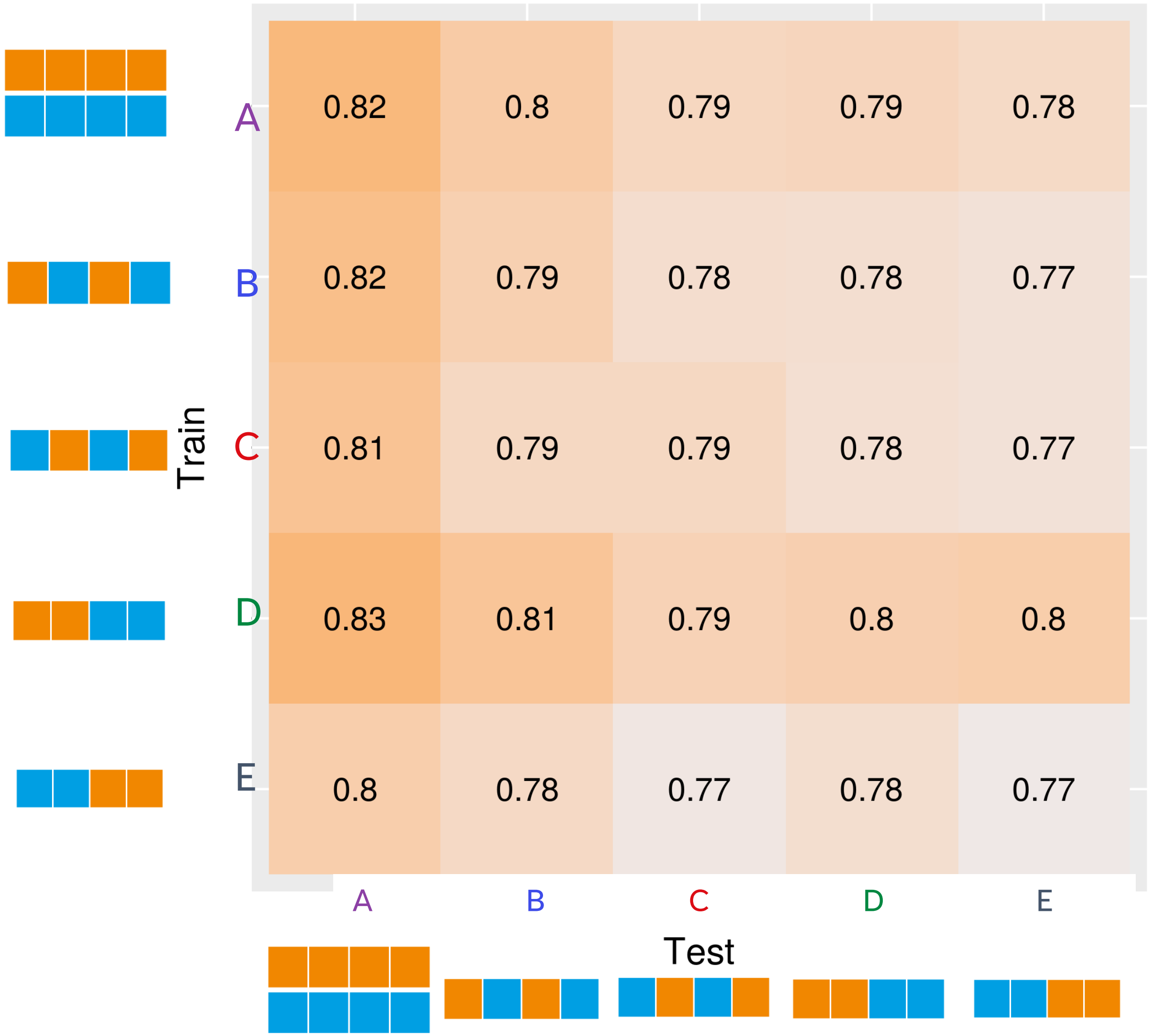


MODEL PERFORMANCE ROC-AUC: BERT

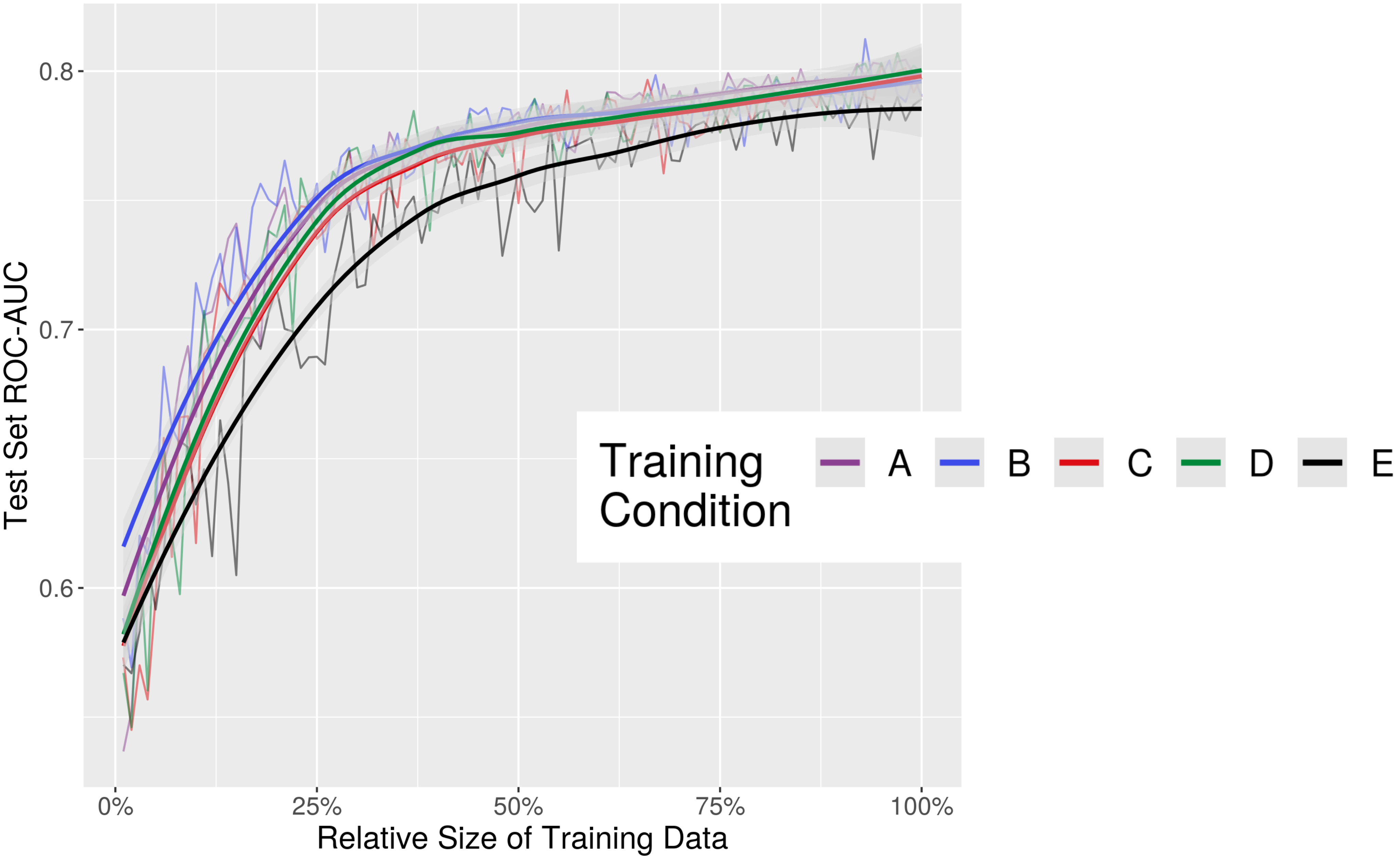
Offensive Language



Hate Speech



LEARNING CURVES: BERT, HATE SPEECH



TAKEAWAYS

- How you collect annotations matters
  - Labels & model accuracy
- Some conditions perform better/worse as train/test data
  - More research needed to inform best practices
- Some evidence of fatigue or motivated misreporting

NEXT STEPS

- Replicate with other training data (images)
- Compare labels from existing annotation platforms
- Vary annotators

THANK YOU

Stephanie Eckman

Social Data Science Center, University of Maryland

steph@umd.edu