CIFellows 2020-2021

Computing Innovation Fellows

Zoey Liu

Department of Computer Science, Boston College

Data-driven Evaluation for Crosslinguistic Low-resource Natural Language Processing

Common model evaluation methods

- focus on largely Indo-European languages
- attend to mostly monolingual settings
- tend to rely on scenarios with large amounts of data
- assume the one data set at hand is ***representative*** of population distribution

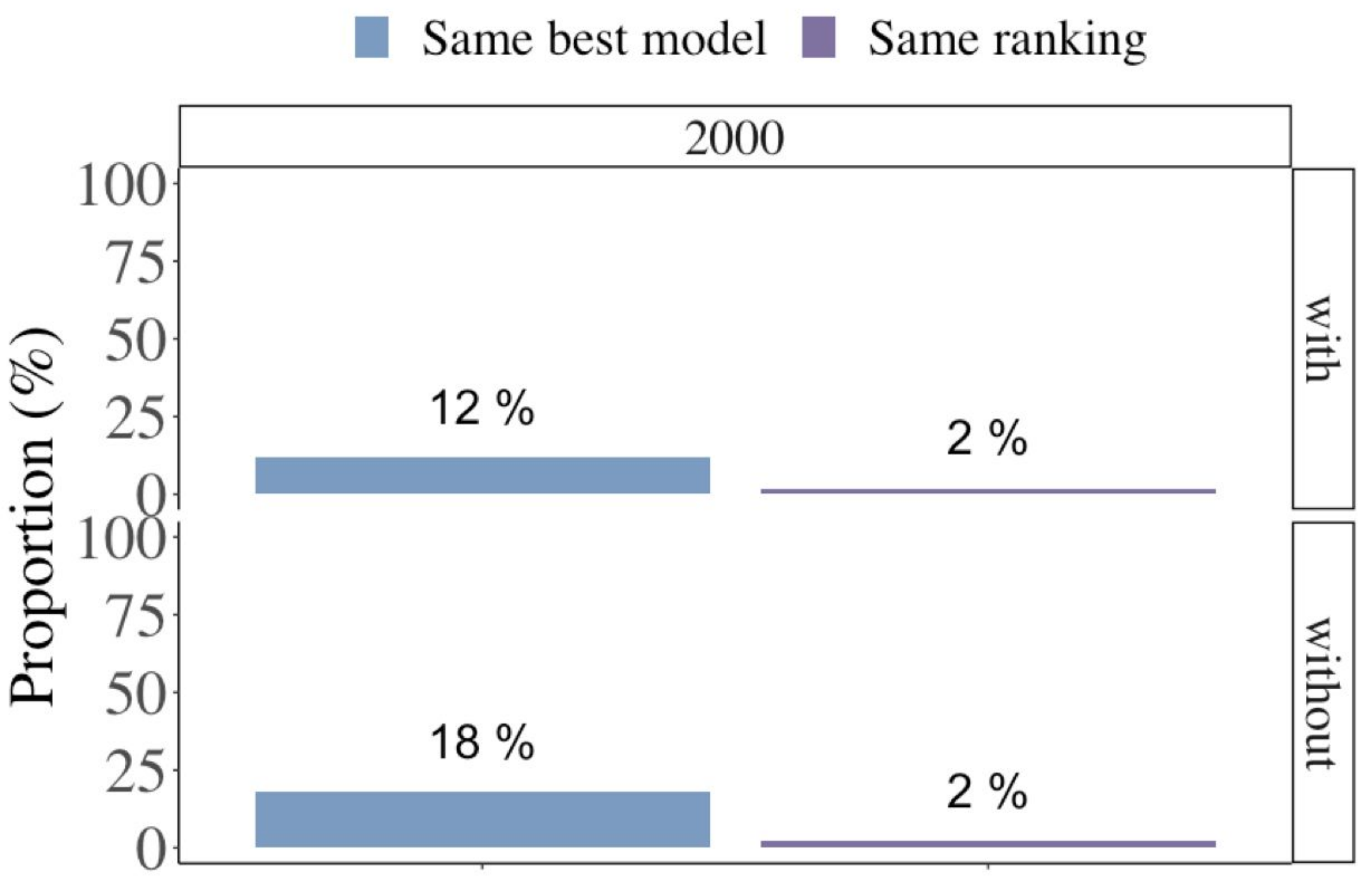
Low-resource settings / Truly low-resource Languages (e.g., endangered languages)

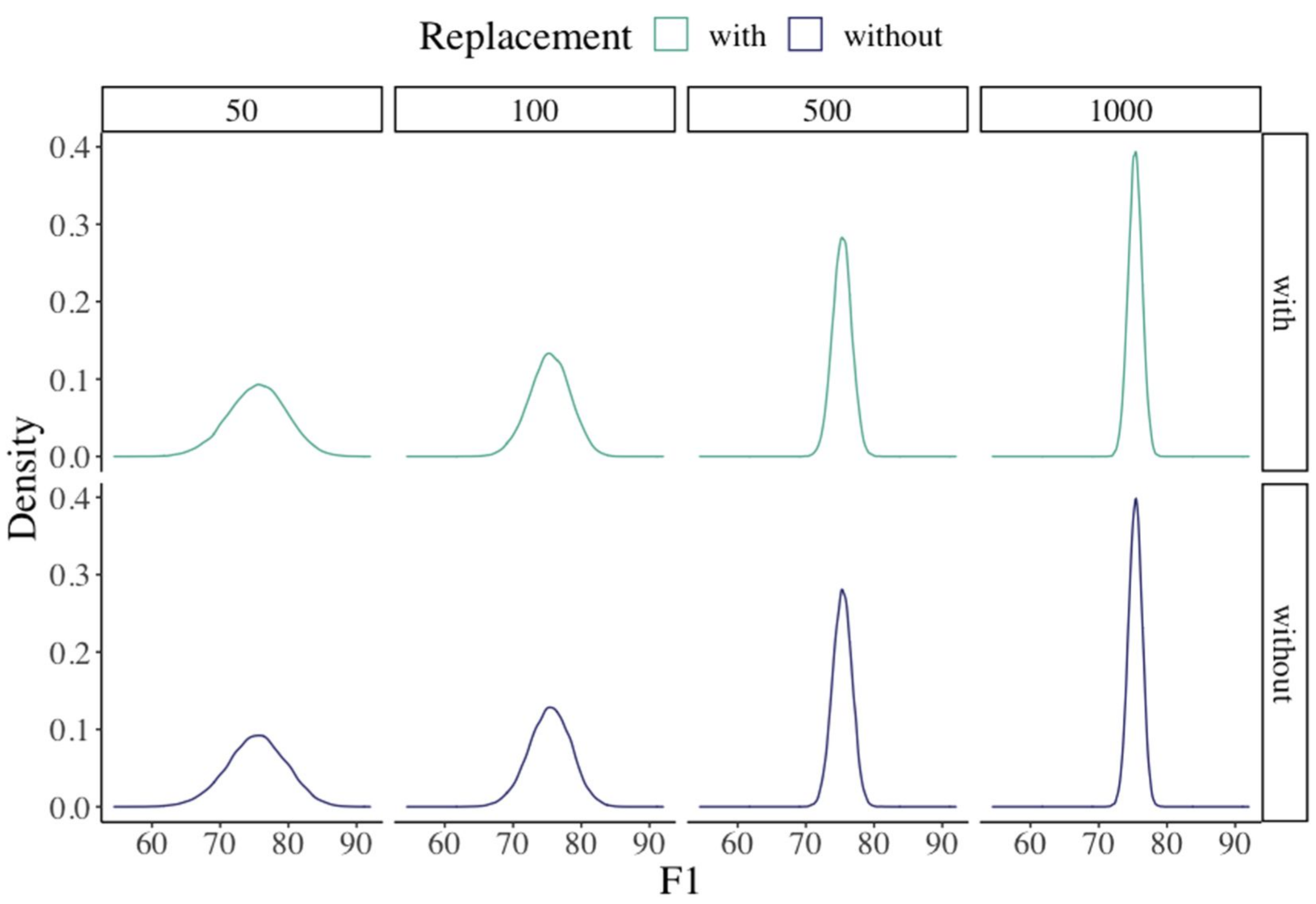
- have only limited data

Therefore assumption about data set representativeness is less likely to hold

Do models generalize to new data sets?

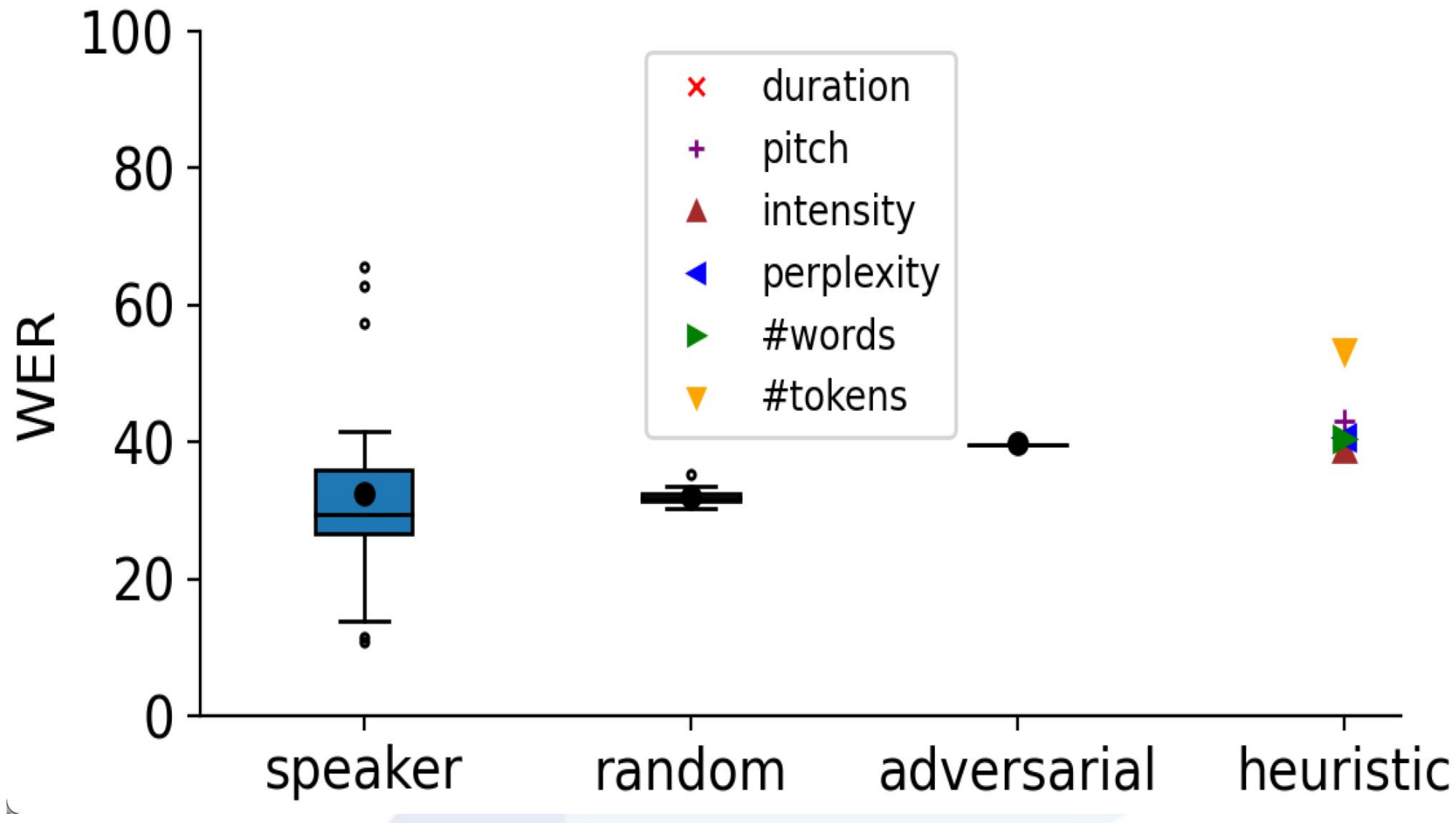
- morphological segmentation as test case
- new randomly sampled data sets of the same size
- new test sets of different sizes
- Models generalize poorly in both augmented low-resource settings and indigenous Mexican languages





How *standard* is standard evaluation?

- automatic speech recognition as test case
- Commonly evaluate models via one pre-defined set of speakers, without cross-validation across speakers
- High variability in acoustic models across speakers for low-resource settings
- For endangered languages, held-out speaker might not be applicable because there is only one speaker in the data



Acknowledgement We are grateful for the continuous support from the Hupa indigenous community.
We would like to especially thank Mrs. Verdena Parker for her generous and valuable input for the documentation work of Hupa throughout the years.

CRA Computing Research Association
CCC Computing Community Consortium Catalyst
NSF