Re-Evaluating the Evaluation of Neural Morphological Inflection Models

Jordan Kodner,[1]
{Salam Khalifa,[1]
Sarah Payne[1]},
Zoey Liu[2]
{first.last}@stonybrook.edu
liu.ying@ufl.edu

ANNs & Morphological Inflection

Applications to Cognitive Science & NLP

- Key role in debates of the **nature of cognitive representations**, renewed by recent advances in **artificial neural networks (ANNs)**
- Standard task in **Natural Language Processing** with **downstream applications**

Mixed Results on Cognitive Feasibility

✅ **Near-ceiling accuracy** on shared tasks in NLP

⚠️ Correlation with **human grammaticality judgments** is mixed

❌ **Learning trajectories & errors** don't match well with humans

Contributions

Creation of **developmentally-plausible data sets** and **robust evaluation techniques** for neural models of morphological inflection

Setup

Data & Evaluation

DATA: three phenomena studied in developmental literature:

- **English past tense:** `CHILDES + UniMorph`, max train = 1000
- **German noun plurals:** `CHILDES + UniMorph`, max train = 600
- **Arabic noun plurals:** `PATB + UniMorph`, max train = 1000

EVALUATION: computational **"wug test"**

- **Train:** given `(lemma, inflected, feature)` triples

```
swim  swam    V; PST
eat   eats    V; PRS; 3; SG
cat   cats    N; PL
```

- **Test:** predict inflected form given `(lemma, feature)` pairs

```
swim  ?   V; PRS; 3;SG     ⇒ swims
box   ?   N;PL             ⇒ boxes
cat   ?   N; SG            ⇒ cat
```

Sampling Strategies

- **UNIFORM:** partition uniformly at random, **5 seeds**
- **WEIGHTED:** frequency-weighted random sampling, **5 seeds**
- **SIGM22:** frequency-weighted random sampling, **1 seed**

Models

- **CHR-TRM** (Wu et al., 2021): a character **transformer**
- **CLUZH** (Wehrli et al., 2022): a character **transducer**
  - **GR** = greedy, **B4** = beam size 4 decoding
- **NONNEUR:** non-neural baseline

Quantitative Analysis

Effect of Training Size

- Weak but significant overall effect **($\beta=0.02$, $p < 0.001$)**
  - **More training ⇒ higher accuracy**
  - **Most significant for CHR-TRM:** sharpest increase in performance
- No significant interaction between **training size & sampling strategy**

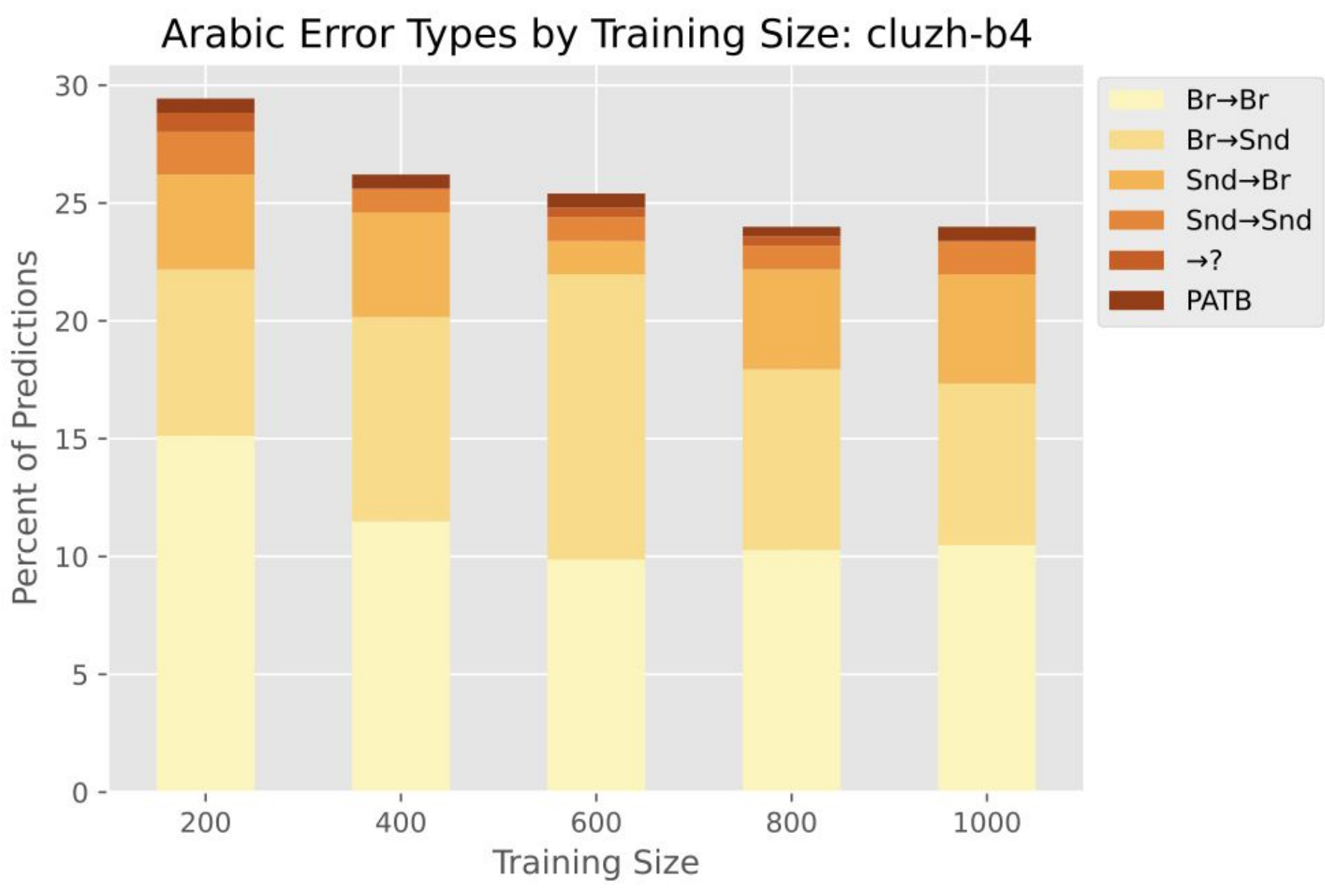
Effect of Sampling Strategy

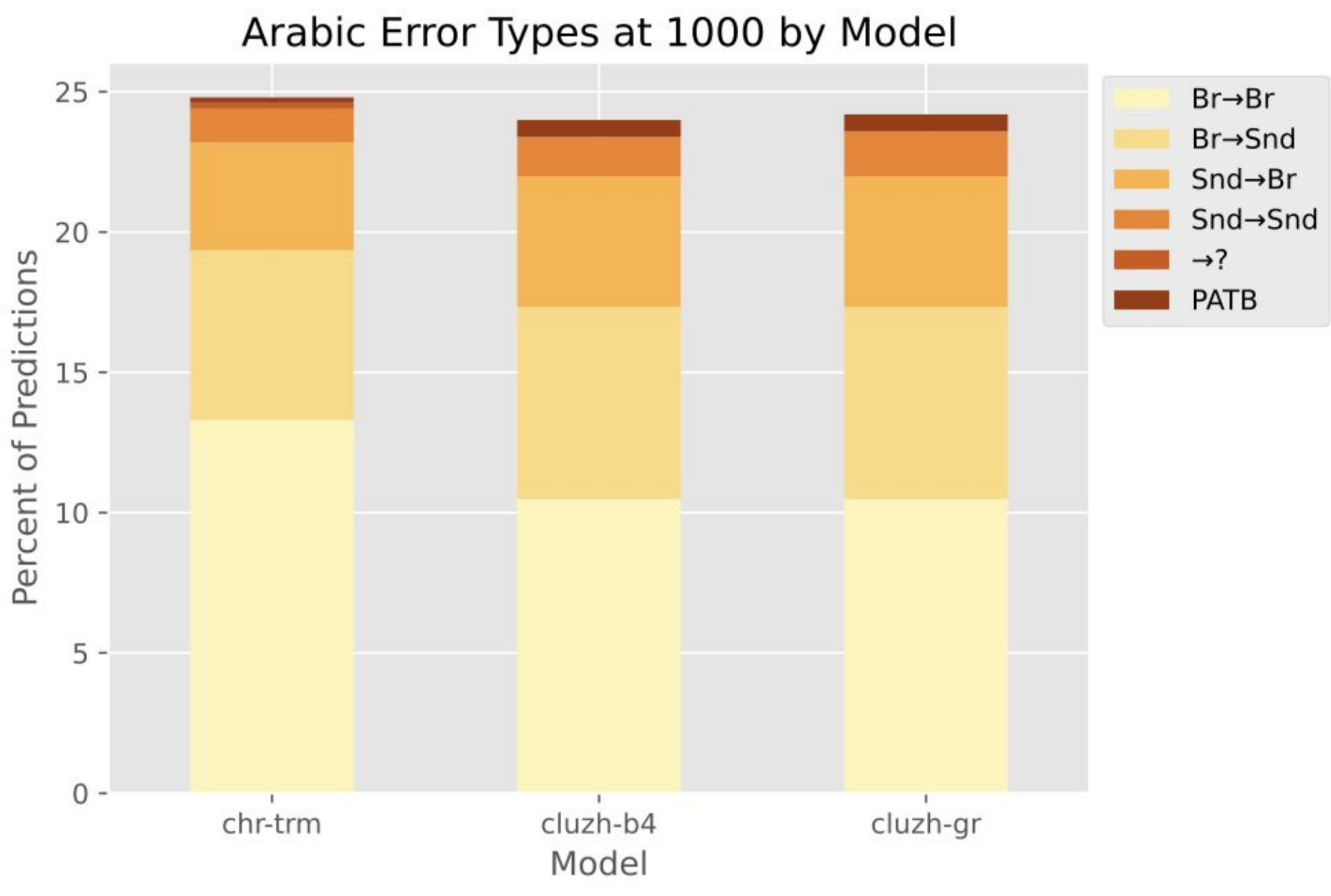
- Higher accuracy for **UNIFORM (67.17%)** than **WEIGHTED (65.24%)**
- Largest effect for **smallest training sizes**
  - **English** (all models) at 100: **66.32% vs. 59.45%**
  - **CHR-TRM** (all languages) at 100: **14.83% vs. 7.42%**
    at 300: **42.69% vs. 30.28%**
- **UNIFORM sampling ⇒ inflated performance**

Variation Across Random Seeds

- Measures of variability:
  - **Score Range:** difference between lowest & highest accuracy
  - **Random seed variability:** standard deviation of accuracy
- **Arabic & German:** higher than English on both measures
- **UNIFORM:** slightly **higher score range** and **comparable random seed variability** to WEIGHTED
- **Training size:** small but significant **negative effect** on both

Arabic Noun Pluralization





Background

- Two types of plurals:
  - **SOUND:** productive suffixation
  - **BROKEN:** unproductive stem mutation
- Relationship between **gender** + suffix
- Two types of **developmental regression:**
  - Overapply FEM sound to **MASC sound & broken**
  - Overapply FEM sound to **MASC & FEM broken**

Results

✅ **BROKEN → SOUND** errors are common

❌ Learning is **monotonic**

- Neither type of **developmental regression**

❌ **BROKEN → BROKEN** errors are **common**

- These are **rare** developmentally

❌ **SOUND → SOUND** errors are **uncommon**

- These are **common** developmentally

❌ **FEM → MASC** errors are relatively **common**

- These are **rare** developmentally

English Past Tense

Background

- **Developmental regression:**
  - Overapply *-ed* to irregulars **(e.g. *goed*)**
- **Over-regularizations** dominate child errors
  - Almost no **over-irregularizations**

Results

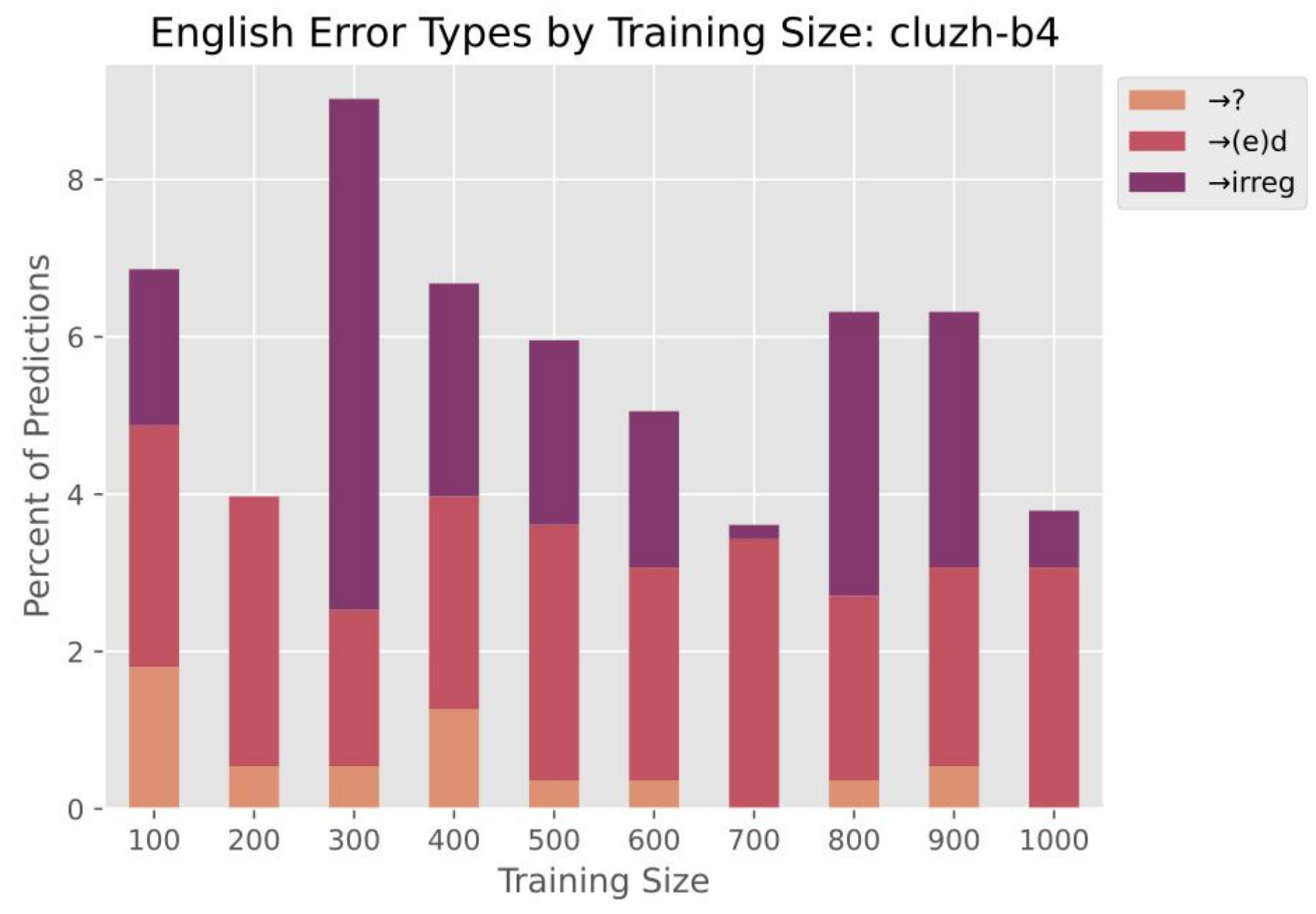
⚠️ CLUZH: **more over-regularizations** than over-irregularizations on full train

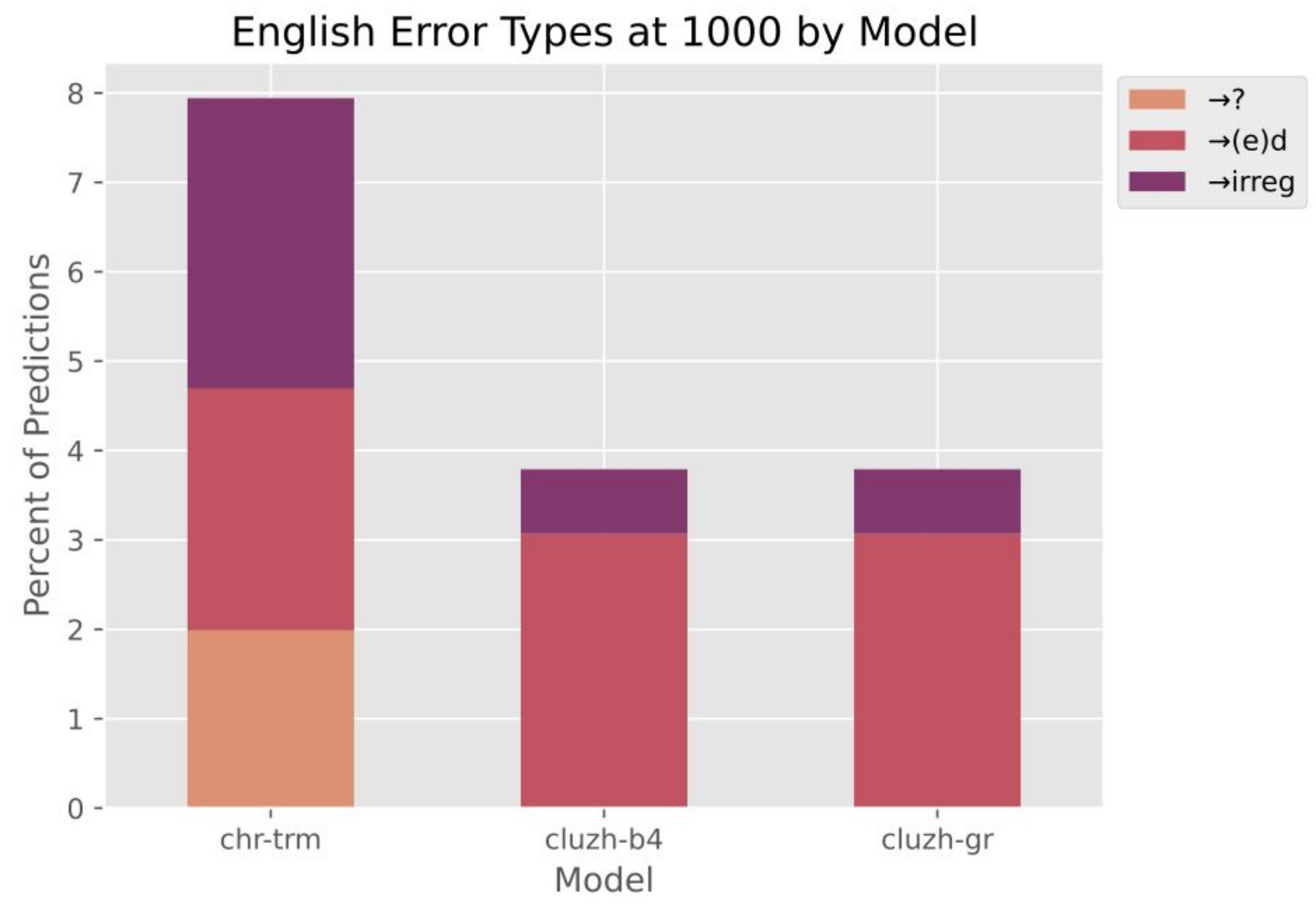
- **Not sufficiently dominant:** order-of-magnitude difference for children

❌ CHR-TRM: **unnatural errors and over-irregularizations** dominate

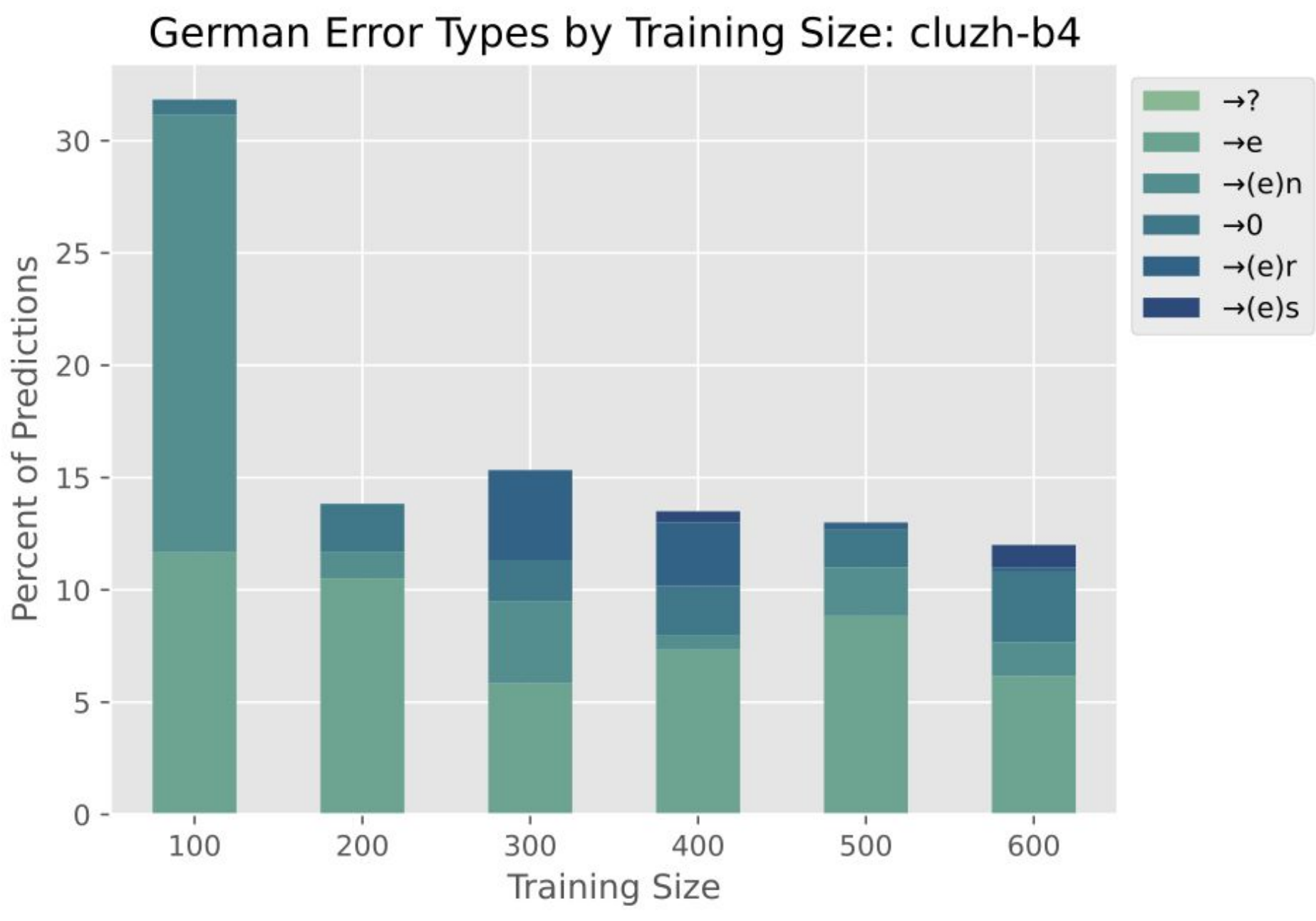
❌ CLUZH-B4: **no developmental regression**

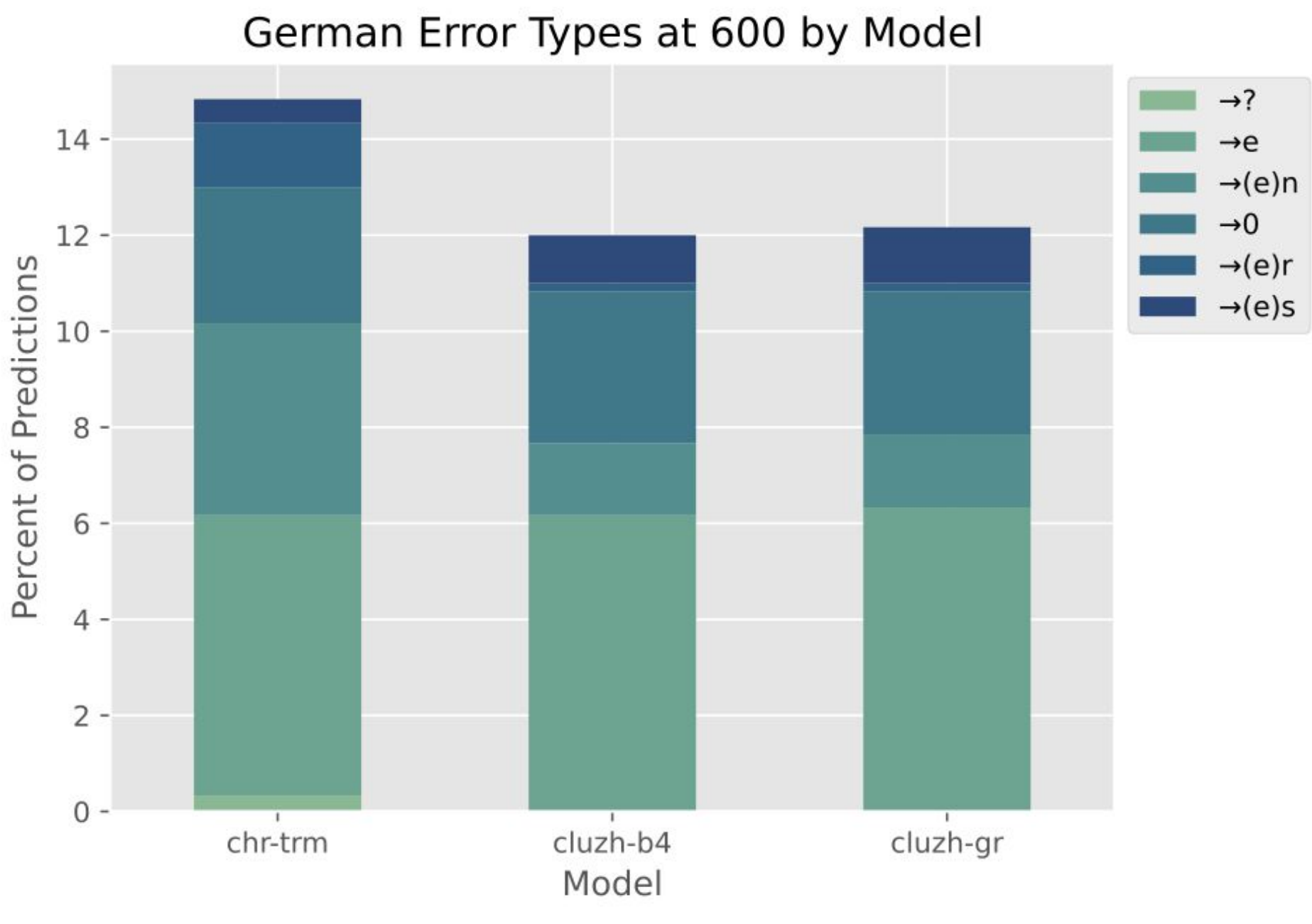
- Error rate & distribution **oscillate**
- **Over-irregularization & unnatural errors** generally too high across sizes
- Error rate spike at 300 = **increase in *over-irregularization***





German Noun Pluralization





Background

- **Five possible processes** for pluralization
- Distinguish **productivity vs. frequency**
  - *-s* = default but **least frequent** (~5%)
  - *-(e)n* = **most frequent**, not default
- No **developmental regression**
  - *-e* and *-∅* acquired **early** & overapplied
  - *-s* acquired **later** & overapplied

Results

✅ **Overapplication of *-e*** at 200 and above

✅ Near-categorical application of ***-(e)n* to FEM**

- *-(e)n* is the **default** FEM affix

✅ Overapplication of *-s* around 300-400

⚠️ Early **dominance of *-(e)n*** at 100

❌ **High overall error rate**