What Drives Performance in Multilingual Language Models?

Sina Bagheri Nezhad and Ameeta Agrawal
PortNLP Research Lab, Portland State University

Factors and Findings

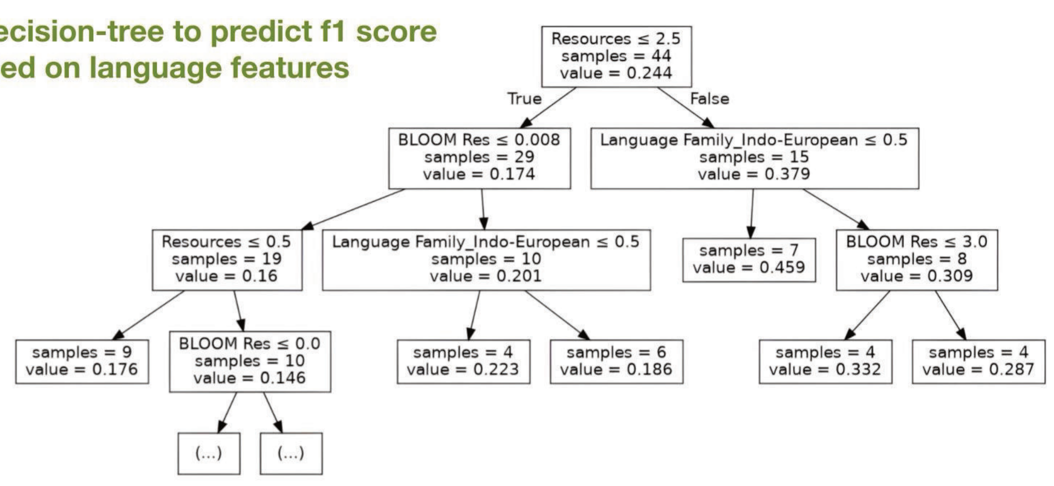
Factors Analyzed

- 📊 **Pretraining Data Size**:
  Crucial for SEEN languages, significantly boosting performance. The most influential factor overall.
- 🏆 **Resource Level**:
  Generally not significant. Only important in models highly correlated with pretraining data.
- 📝 **Language Family**:
  Key for UNSEEN languages. Models rely on linguistic relationships for cross-lingual transfer. Some models indicate a predisposition for particular language families, such as Indo-European.
- ✍️ **Script Type**:
  Crucial for UNSEEN languages. Models depend on script similarities to generalize to new languages. Specific models show preferences for certain script types, such as Latin and Devanagari.

Additional Explorations

- 🧩 **Model Size and Architecture**:
  Do not significantly alter the importance of pretraining data size, language family, and script type.
- 🧪 **Training Scenarios**:
  Zero-shot, two-shot ICL, and full-shot fine-tuning do not significantly change the importance of key factors.
- 📊 **Presence or Absence**:
  For some models, the impact of pretraining data size actually indicates the presence or absence of a language rather than the amount of data.
- 📚 **Pretraining and Instruction-tuning**:
  Pretraining data distribution is more crucial than fine-tuning data for instruction-tuned models. Initial pretraining plays a critical role.

Experiment Process



2) Train decision-tree to predict f1 score based on language features

Language Name	Script	Language Family	ResLevel	% Pretrained data	F1-score Zero-shot Bloom-560m
⋮	⋮	⋮	⋮	⋮	⋮
Ukrainian	Cyrl	Indo-European	3	0	0.170
Umbundu	Latn	Niger-Congo	0	0	0.173
Urdu	Arab	Indo-European	3	0.1	0.123
Northern Uzbek	Latn	Turkic	3	0	0.219
Venetian	Latn	Indo-European	1	0	0.347
Vietnamese	Latn	Austro-Asiatic	4	2.7	0.354
Waray-Waray	Latn	Austronesian	0	0	0.258
Wolof	Latn	Niger-Congo	2	0.004	0.303
Xhosa	Latn	Niger-Congo	2	0.001	0.145
Eastern Yiddish	Hebr	Indo-European	1	0	0.098
Yoruba	Latn	Niger-Congo	2	0.006	0.164
⋮	⋮	⋮	⋮		⋮

1) Run model in a specific setup and evaluate results by f1-score

Null hypothesis (H0): There is no significant difference in f1-score between languages with a resource level less than or equal to 2.5 and those with a resource level greater than 2.5.

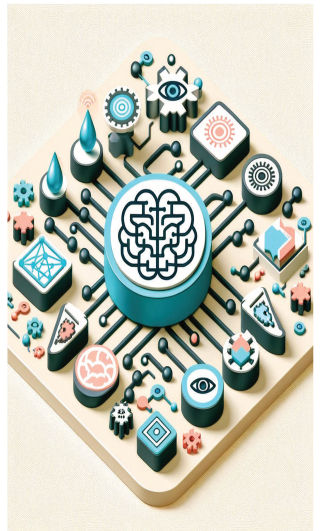
Alternative hypothesis (H1): There is a significant difference in f1-score between languages with a resource level less than or equal to 2.5 and those with a resource level greater than 2.5.

3) Formulated null and alternative hypotheses.

4) Run Mann-Whitney U test to test hypothesis.

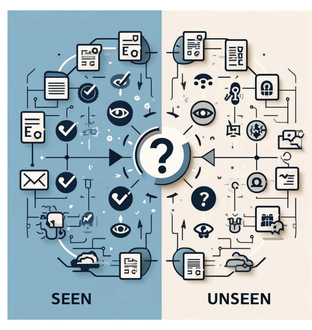
P-value < 0.001
Null hypothesis rejected.

Methodology



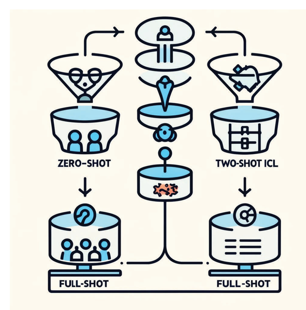
Models
- mBERT
- XLM-R
- GPT-3.5
- BLOOM in 5 sizes
- XGLM in 4 sizes
- BLOOMZ in 5 sizes

SIB-200 Dataset
- 204 languages
- 21 language families
- 29 script types



Seen VS Unseen
- ALL languages
- SEEN languages
- UNSEEN languages

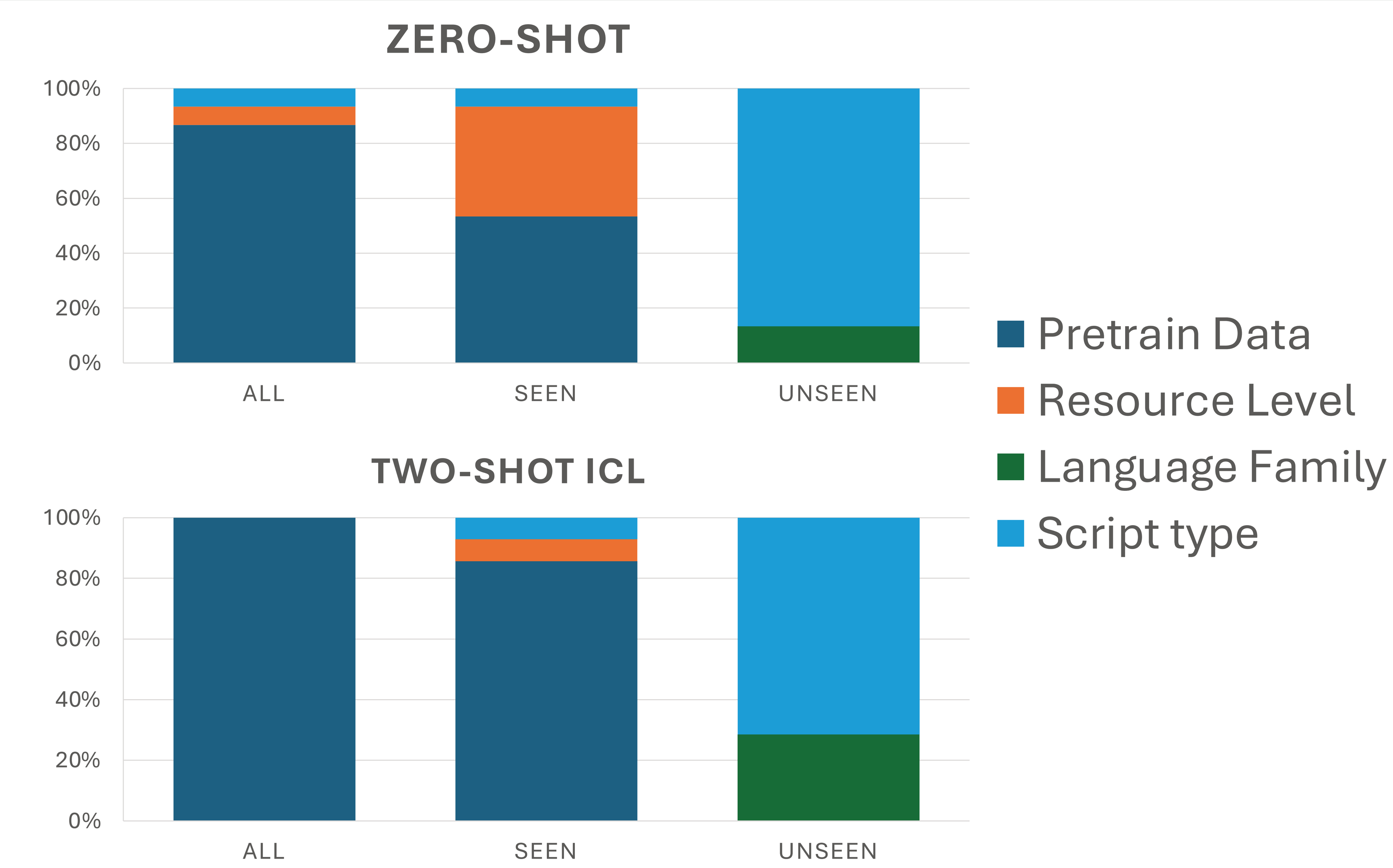
Scenarios
- Zero-shot
- Two-shot ICL
- Full-shot.



Analysis Method

- 🌍 **Extensive Evaluations**: Conducted evaluations across 204 languages in 93 different model and scenario combinations.
- 🌳 **Decision Tree Analysis**: Created 93 decision trees to determine the importance of different factors.
- 🔍 **Statistical Testing**: Used the Mann-Whitney U test to validate the significance of identified features.

Results



Acknowledgements

This work was supported by the National Science Foundation through grants CRII:RI 2246174 and SAI-P 2228783.