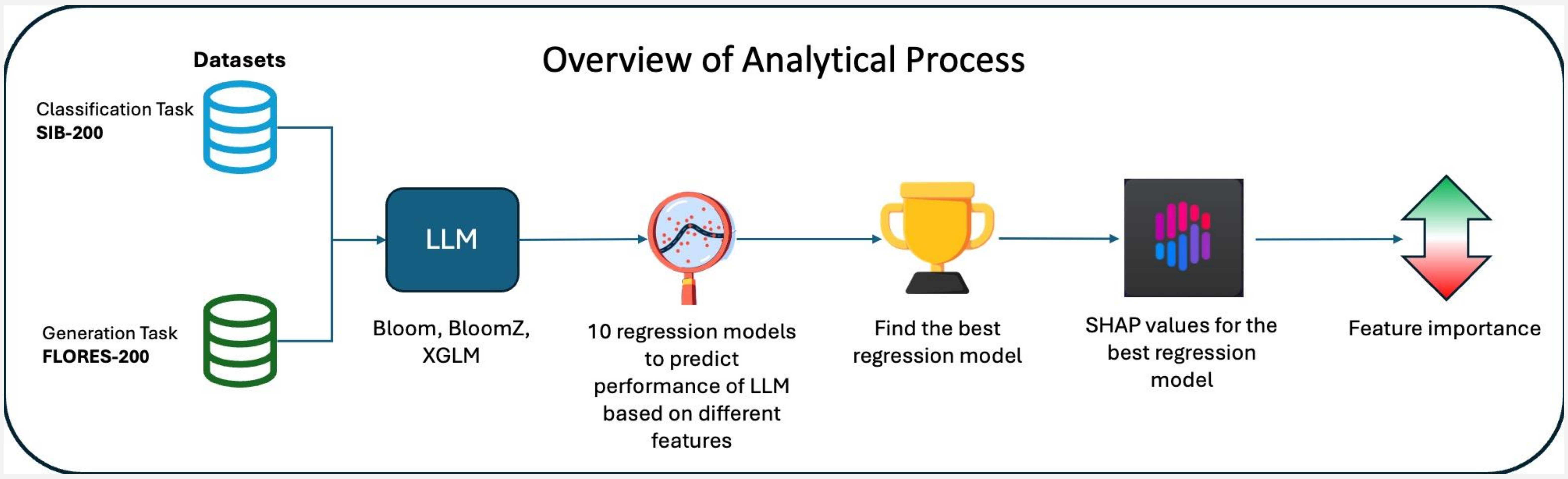
Beyond Data Quantity: Key Factors Driving Performance in Multilingual Language Models

Sina Bagheri Nezhad, Ameeta Agrawal, Rhitabrat Pokharel
Department of Computer Science
Portland State University, USA

Introduction

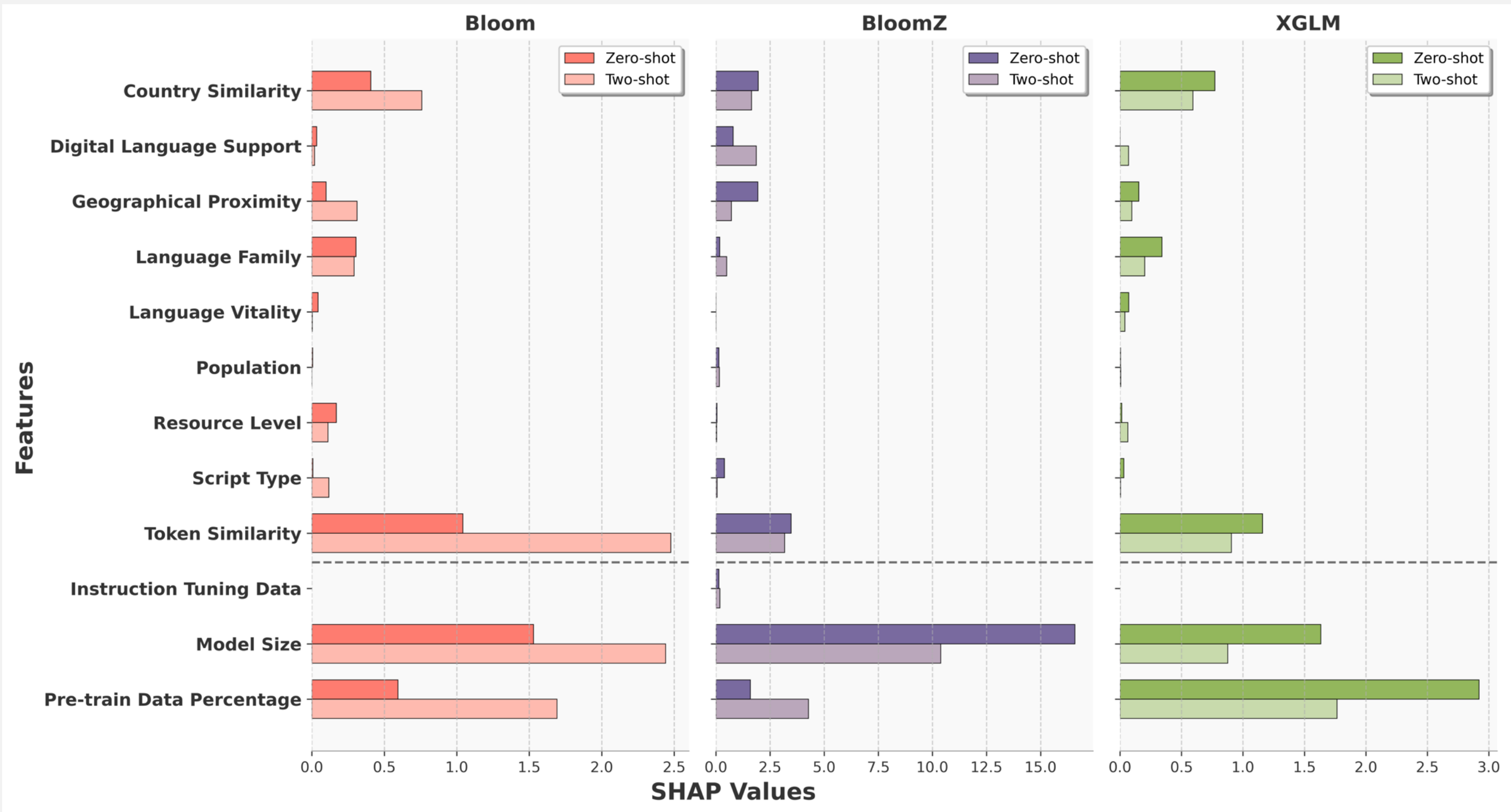
- Multilingual Language Models (MLLMs) show performance disparities across languages.
- We explore factors beyond pre-train data size and model size that influence MLLM performance.
- Analyzed 204 languages using **SIB-200 (classification)** and **Flores-200 (machine translation)** datasets.

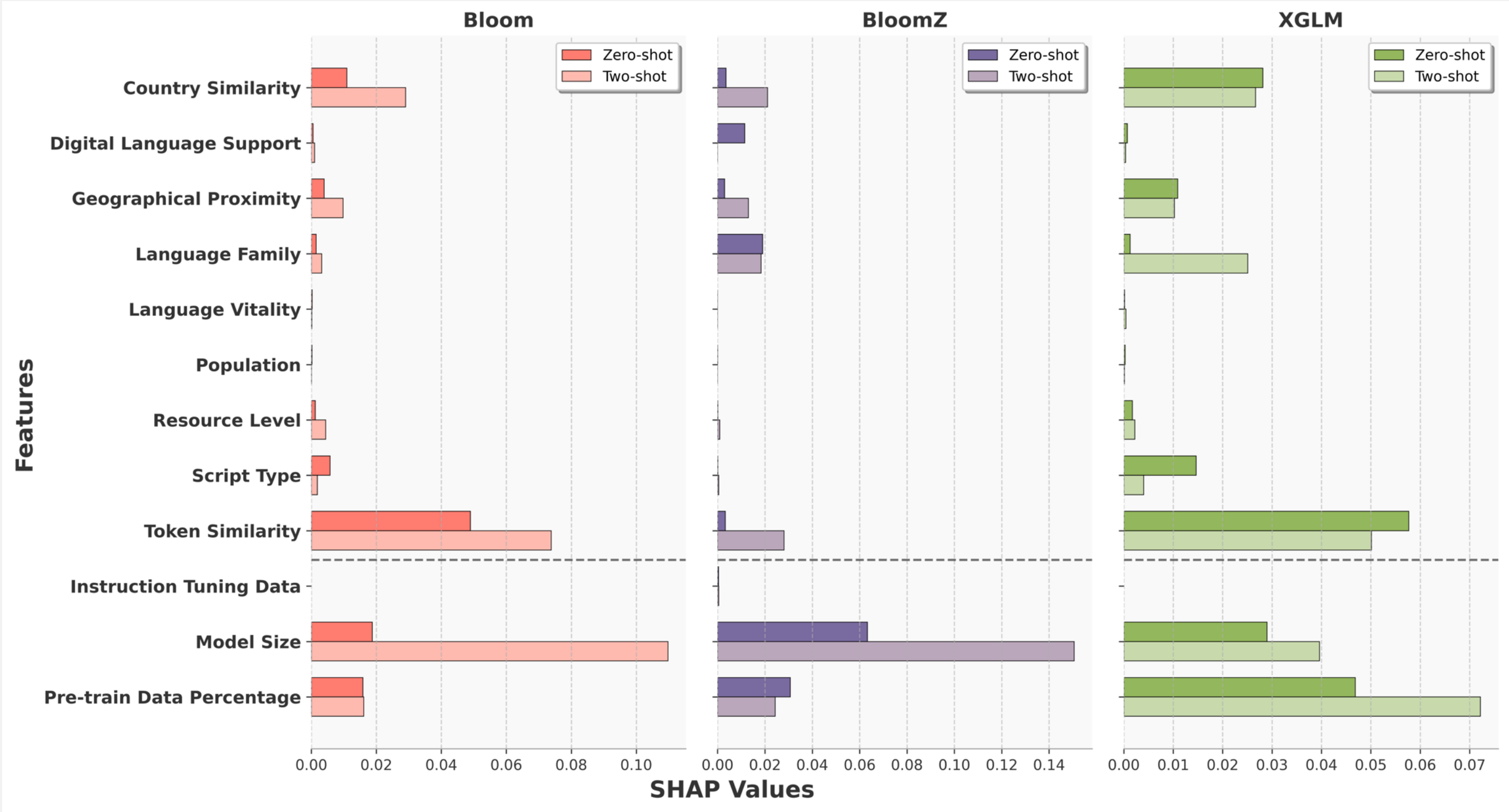

Overview of Analytical Process
Datasets
Classification Task
SIB-200
Generation Task
FLORES-200
LLM
Bloom, BloomZ, XGLM
10 regression models to predict performance of LLM based on different features
Find the best regression model
SHAP values for the best regression model
Feature importance

Methodology

- **Models Evaluated**: Bloom, BloomZ, XGLM (various sizes)
- **Tasks**: Classification & Generation in Zero-shot & Two-shot settings
- **Key Features Analyzed**:
  - Model-specific: Model size, Pre-train data, Instruction tuning (BloomZ)
  - Language-specific: Script type, Token similarity, Language family
  - Socio-linguistic: Population, Language vitality, Digital support
  - Geographical & Cultural: Geographical proximity, Country similarity

Results





Key Findings

Feature Importance (via SHAP Analysis)

- **Token Similarity & Country Similarity** emerged as pivotal factors.
- **Pre-train data percentage & Model size** significantly impact performance.
- **Geographical proximity had minimal effect**; country similarity was more influential.
- **Instruction tuning (BloomZ) had little impact compared to pre-train data.**

Regression Model Performance

- **Ensemble models (XGBoost, Random Forest, Gradient Boosting) performed best** in predicting MLLM success.
- **Linear models struggled**, highlighting non-linearity in multilingual modeling.

Takeaway

- Token similarity enhances cross-lingual transfer.
- Country similarity plays a crucial role in language model effectiveness.
- Pre-train data remains the most critical factor.
- Insights can guide future development of **more equitable and effective** multilingual models.

Scan me

Acknowledgments

This work was supported by the National Artificial Intelligence Research Resource (NAIRR) Pilot, funded by the National Science Foundation under award No. 240158.