HumanEval-XL: A Multilingual Code Generation Benchmark for Cross-lingual Natural Language Generalization

Qiwei Peng*, Yekun Chai*, Xuhong Li

UNIVERSITY OF COPENHAGEN

Baidu 百度

qipe@di.ku.dk chaiyekun@gmail.com

Code Generation

For example:

```
def binomial_coeff(n,k):
    """ Write a Python function to
find binomial coefficient.
    >>> binomial_coeff(5,2)
    10
    """
```

Given input

Model

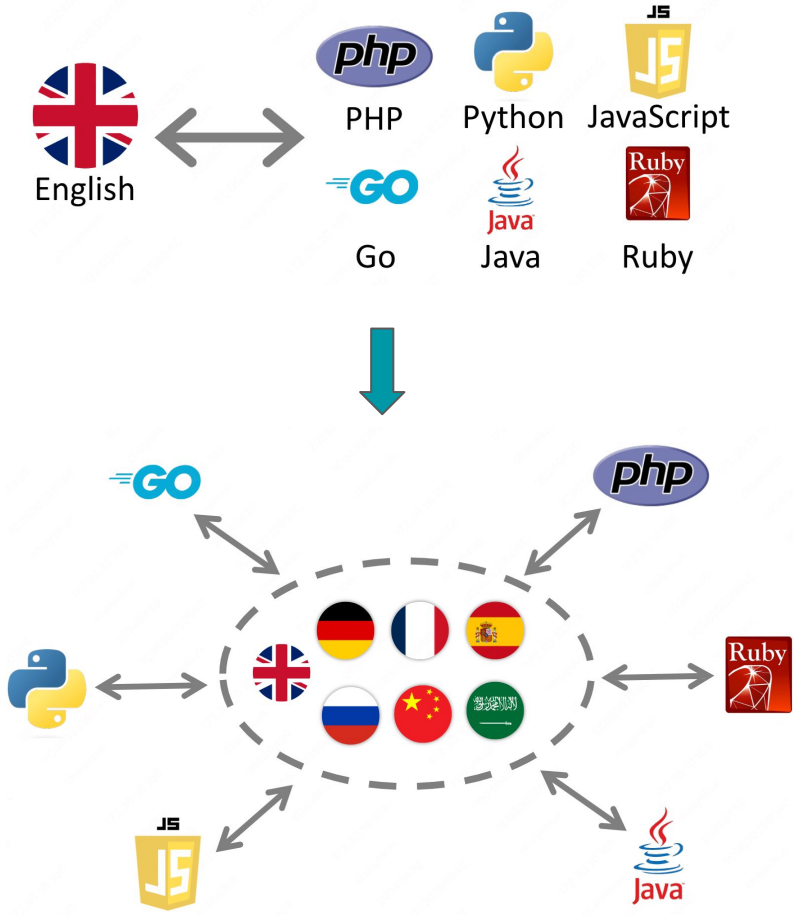
```
    if k > n :
        return 0
    if k==0 or k ==n :
        return 1
    return binomial_coeff(n-1,k-1)
+ binomial_coeff(n-1,k)
```

Expected output

The task can be formulated in different forms (e.g., code completion, variable/line infilling).

Motivation

- ➔ Current benchmarks primarily focus on **English** for code generation, limiting the relevant evaluation of LLMs on cross-lingual transfer.
- ➔ **High quality cross-lingual (NL) code generation benchmark** helps building better code generation models, leading to advanced code applications of **global impact and easy access for people from different regions**.

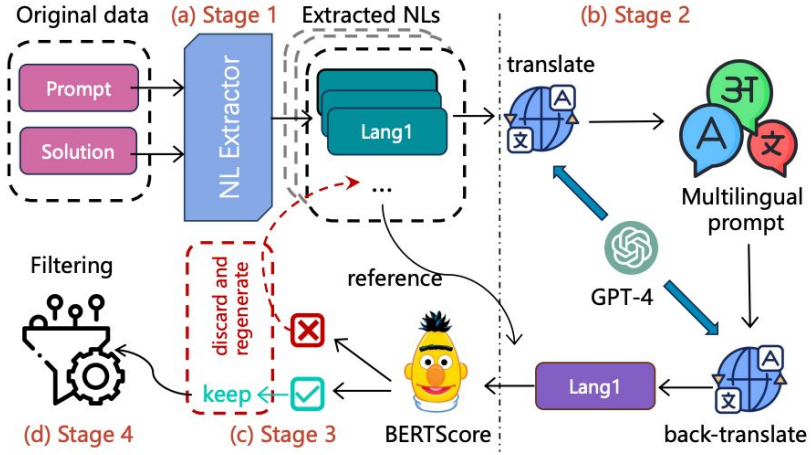


Challenge

- ★ **Complexity of Task**: Code generation involves understanding complex instructions and translating them into syntactically correct code—a challenge compounded when adding multiple languages.
- ★ **Quality Assurance**: Ensuring the translations maintain semantic integrity across languages is critical for the model's training and performance evaluation.
- ★ **Consistent Evaluation**: Developing a fair and unbiased evaluation metric across multiple languages and programming languages is crucial for true performance assessment.

Methods

A. Text Extraction (**Stage 1**): We extract NL texts from the prompt.
B. Translation and Back-Translation (**Stage 2**): The extracted texts are translated into 23 different languages using GPT-4. These translations are then back-translated to English for subsequent automatic quality checks.
C. Quality Assessment with BERTScore (**Stage 3**): Stage 3 assesses translation quality by computing the BERTScore similarity score between the original English text and its back-translated text. Translations with a low similarity score (threshold < 0.95) are rejected and subjected to re-translation (max # of iter = 3).
D. Quality Control (**Stage 4**): Heuristic checks and manual evaluations are performed on the quality of the translated texts.



Our dataset was constructed through a rigorous four-stage process including extraction of natural language prompts, translation to 23 languages, back-translation for quality check, and quality control using BERTScore.

Our Dataset: HumanEval-XL

Dataset	#Samples	#Average Test Cases	Data source	#PL	#NL	Parallel?
HumanEval (Chen et al., 2021)	164	7.7	Hand-written	1	1	✗
MBPP (Austin et al., 2021)	974	3.0	Hand-written	1	1	✗
APPS (Hendrycks et al., 2021)	10,000	13.2	Competitions	1	1	✗
DSP (Chandel et al., 2022)	1,119	2.1	Github Notebooks	1	1	✗
MTPB (Nijkamp et al., 2023b)	115	5.0	Hand-written	1	1	✗
DS-1000 (Lai et al., 2023)	1,000	1.6	StackOverflow	1	1	✗
Multlingual HumanEval (Athiwaratkun et al., 2023)	1,935	7.8	Hand-written	12	1	✗
ODEX (Wang et al., 2022)	945	1.8	StackOverflow	1	4	✗
HumanEval-XL	**22,080**	8.3	Hand-written	**12**	**23**	✔

HumanEval-XL consists of 80 parallel coding problems spanning 12 PLs and 23 NLs. In total, this benchmark includes 22,080 coding problems.

Family	Languages
Afro-Asiatic	Arabic, Hebrew
Austro-Asiatic	Vietnamese
Austronesian	Indonesian, Malay, Tagalog
Indo-European (Germanic)	English, Dutch, German, Afrikaans
Indo-European (Romance)	Portuguese, Spanish, French, Italian
Indo-European (Greek)	Greek
Indo-European (Iranian)	Persian
Slavic	Russian, Bulgarian
Sino-Tibetan	Chinese
Turkic	Turkish
Uralic	Estonian, Finnish, Hungarian

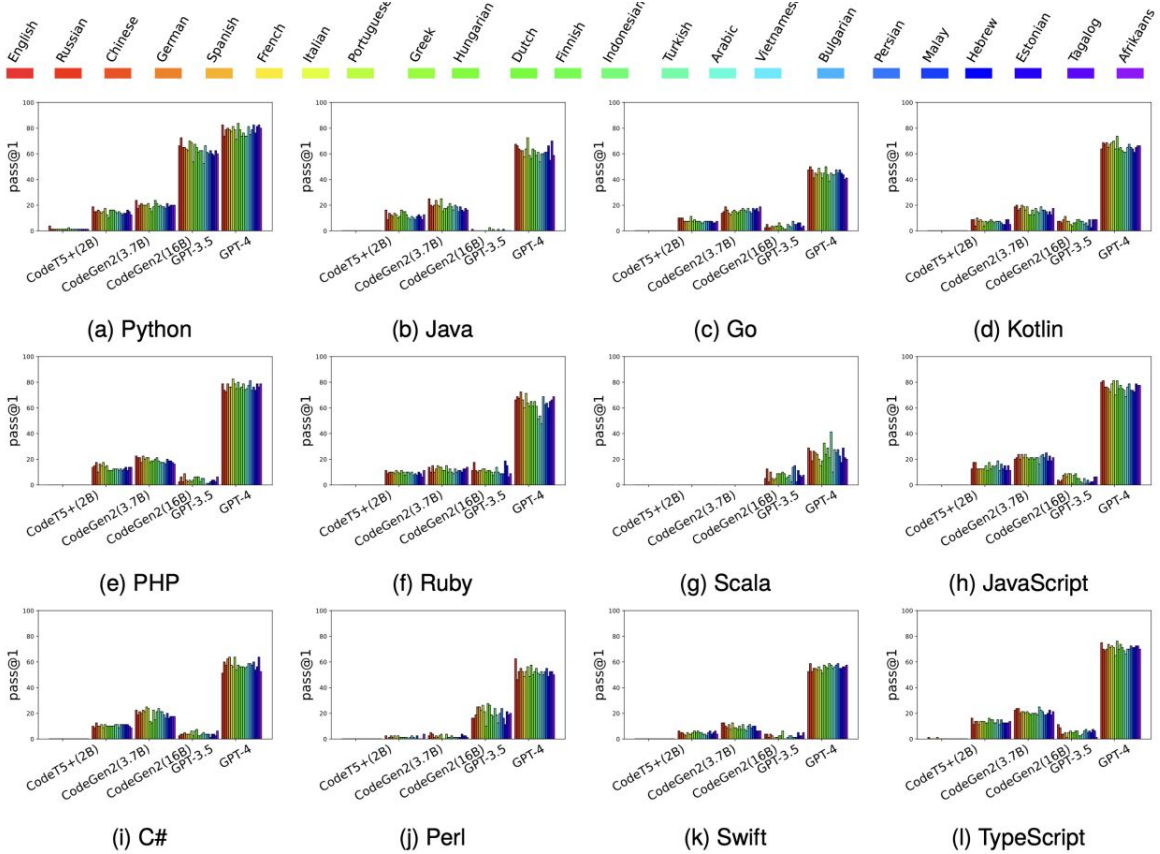
It spans across 11 distinct language families.

Experiments

➔ We tested different models including CodeT5+, CodeGen2, GPT-3.5, and GPT-4.

Findings:

- Clear cross-lingual inconsistency.
- Increase in model size boosts performance.
- Specialized code pre-training plays a pivotal role in code generation.



Key Observations

- Despite advancements, LLMs struggle to maintain semantic consistency across different languages in the context of code generation.

- Increased model size correlates with improved performance, suggesting that scaling model parameters is beneficial for multilingual understanding.

Conclusion

- We propose **HumanEval-XL**, a massively multilingual code generation benchmark for assessing cross-lingual NL generation for LLMs.
- Our study reveals the **inconsistent cross-lingual transfer** of current LLMs (code/general), underscoring the significant challenge in achieving effective cross-lingual NL generalization.