INCLUDE BENCHMARK

Evaluating **Multilingual** Language Understanding with **Regional Knowledge**

Angelika Romanou • **Negar Foroutan** • Anna Sotnikova • Zeming Chen • Sree Harsha Nelaturu • Shivalika Singh • Rishabh Maheshwary • Micol Altomare • Mohamed A. Haggag • Imanol Schlag • Marzieh Fadaee • Sara Hooker • Antoine Bosselut

Missing your language?

Join us for INCLUDE-v2

Multilingual benchmarks should reflect the regional and cultural knowledge of where languages are spoken or used, not be translated from western-centric resources

The **largest multilingual exam collection** to date with ~200K question-answer pairs focusing on regional & cultural knowledge

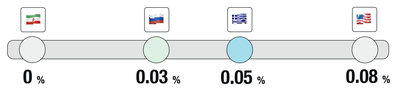
Limitations of the current translated benchmarks

- They primarily reflect US / Western-centric knowledge
- They may contain errors introduced during translation
- They often exhibit translationese artifacts

Motivation: The same questions can have different answers depending on where they are asked

REGIONAL KNOWLEDGE — Law & Regulations

Ποιο είναι το επιτρεπτό όριο αλκοόλ ανά λίτρο αίματος στην οδήγηση ;	Какой уровень алкоголя в крови допустим при вождении?
What is the Blood Alcohol Limit (BAC%)for driving?	What is the Blood Alcohol Limit (BAC%)for driving?



A balanced subset is available on Hugging Face

CohereForAI/include-base-44

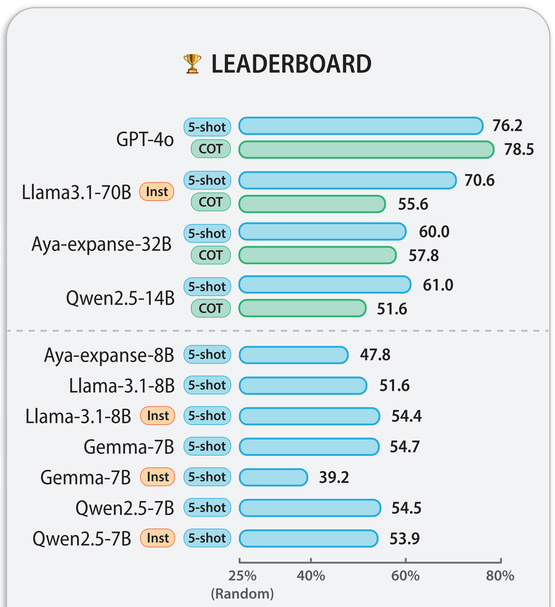
Tasks: Multiple Choice | Languages: 44 Languages | Size: 22,637 | ArXiv: arxiv: 2411.19799

44 LANGUAGES

Azerbaijani	Bulgarian	Greek	Armenian	Basque
Croatian	Hungarian	Nepali	Macedonian	Tagalog
Serbian	Albanian	Lithuanian	Malayalam	Georgian
Bengali	Estonian	Turkish	Belarusian	Telugu
Dutch	Hebrew	Hindi	Malay	Urdu
Kazakh	Arabic	Korean	Ukrainian	Tamil
Uzbek	Russian	Spanish	Chinese	Italian
French	German	Finnish	Persian	Indonesian
Vietnamese	Portuguese	Japanese	Polish	

57 TASKS: Arts & Humanities | Social Sciences | STEM | Business & Commerce | Health education | Professional Licenses | Occupational Licenses | +

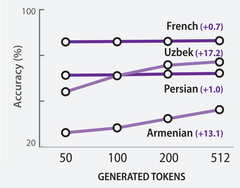
HOW DO LLMs PERFORM ON INCLUDE ?



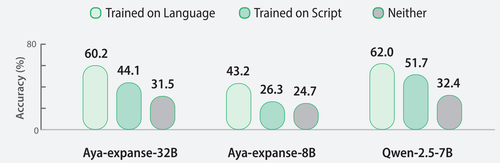
Multilingual LLMs **do not follow instructions** the same way in all languages.

GPT-4o shows **3.1%** gain when **increasing the generation length window**

Perfromance on GPT-4o with different generation windows on **high** vs **low** resource languages



Open models **transfer ~20%** more **to unseen languages with shared scripts** than to those with different ones



Increased model size **significantly enhances multilingual capabilities** (with the same pre-training data)

Instruction tuning may hurt the multilingual ability likely due to **English-heavy post-training**

Models **struggle** more on tasks requiring **regional knowledge** than those assessing universal knowledge

English prompts offer **modest improvements** of **1-2%** compared to **In-language prompts**

DATA CONTRIBUTORS

Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang

EPFL

nlp

Cohere For AI

@agromanou.bsky.social
@negarforoutan.bsky.social

@agromanou
@negarforoutan

angelika.romanou@epfl.ch