Lost in the **Tower of Babel**: The Adverse Effects of **Incidental Multilingualism** in LLMs

!! HIRE ME (Postdoc) !!



Anjishnu

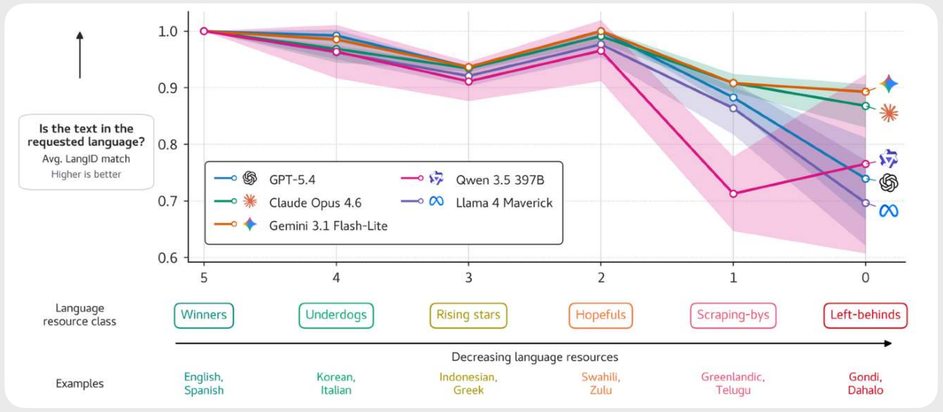


Chutong



Antonis

GEORGE MASON UNIVERSITY



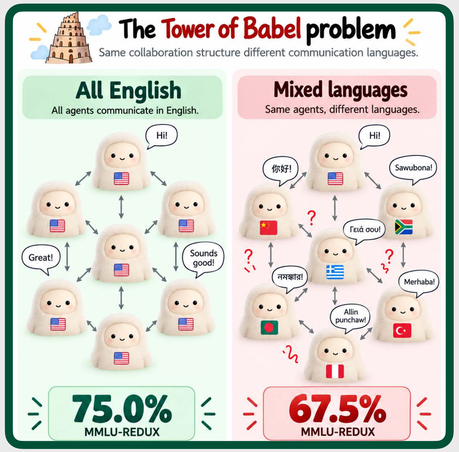


How to make an LLM multilingual?

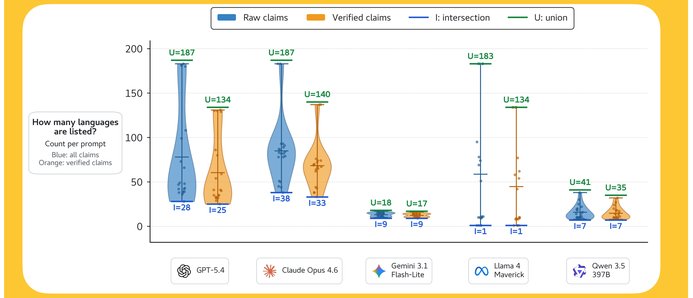
Surprising answer

Speech and Language Processing (3rd ed. draft)
Dan Jurafsky and James H. Martin

Just throw all the languages data in a big file and pretrain on it!



RESULTS



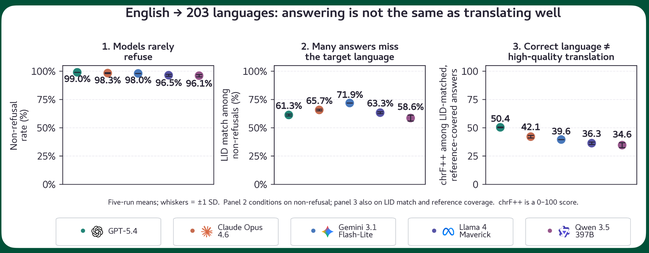
1. The lowest resource languages create a clear separation between model providers.

2. Model providers **SHOULD** publish language-level evaluation cards.

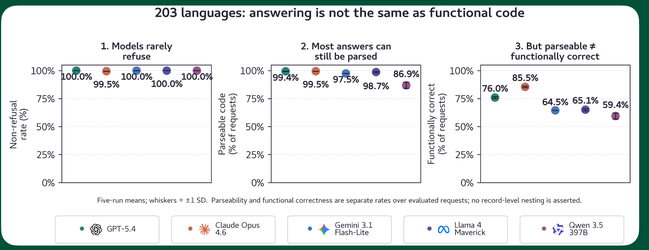
3. Multilingualism **SHOULD** be a training objective, not incidental.

More RESULTS

Machine Translation



Code Generation



4. **Refusals** >> low-quality/incorrect responses !!

***Even more* RESULTS** →