Don’t Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs

Xiang Zhang*, Senyu Li*, Bradley Hauer, Ning Shi, Grzegorz Kondrak

Alberta Machine Intelligence Institute, Dept of Computing Science
University of Alberta, Edmonton, Canada

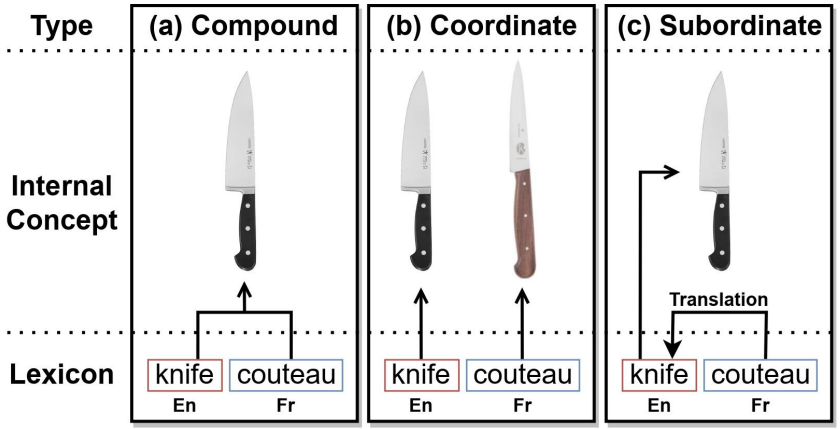
“*” indicates equal contribution

Three Types of Bilingualism

a) **Compound**
Languages are equally dominant, with a single mental representation.

b) **Coordinate**
Separate mental representations for each language

c) **Subordinate**
Input is first translated into the dominant language, before answers are formulated.

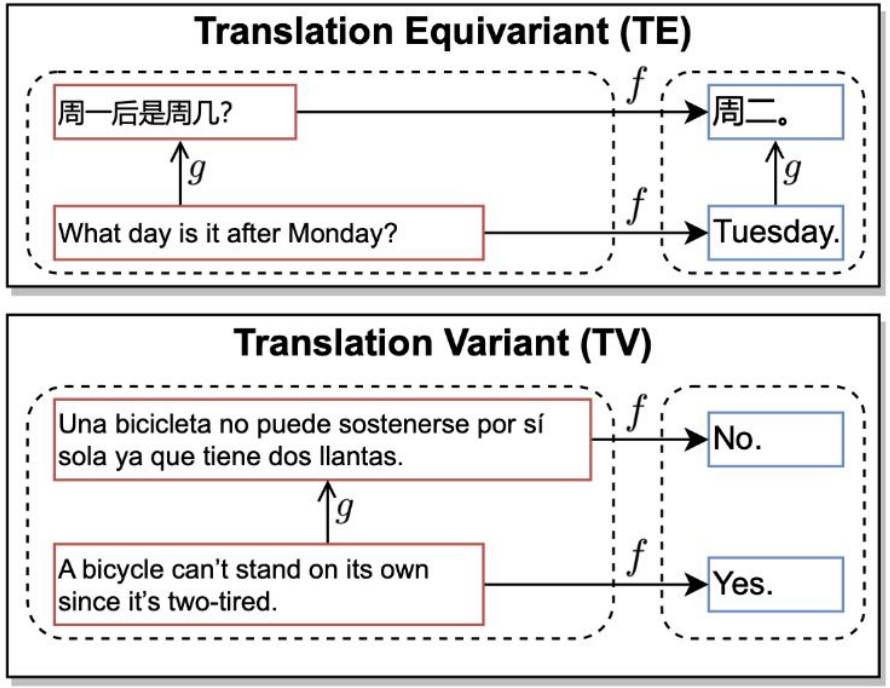


Three Types of Bilingualism (D'Acierno, 1990).

Translation Equivariant vs. Translation Variant

- Translation Equivariant tasks: Translating the question **does not** change the answer.

- Translation Variant tasks: Translating the question **may** change the answer.



Translation Equivariant Tasks Tested

- Math Reasoning
  Example: Ten more than twice the number of birds on the fence is 50. How many birds are on the fence?

- Knowledge Access
  Example: Who created the character of sherlock holmes?
  A. Arthur Conan Doyle B. Eugen Bauder C. Patrick Ribbsaeter D. Taylor Lautner E. Peter Wentz

- Common Sense Reasoning
  Example: John looked for beavers but couldn't find any, because he lived where?
  A. America B. Australia C. Countryside D. Dictionary E. Woodlands

Translation Variant Tasks Tested

- Pun detection
  Example Pun: A bicycle can't stand on its own because it is two-tired.

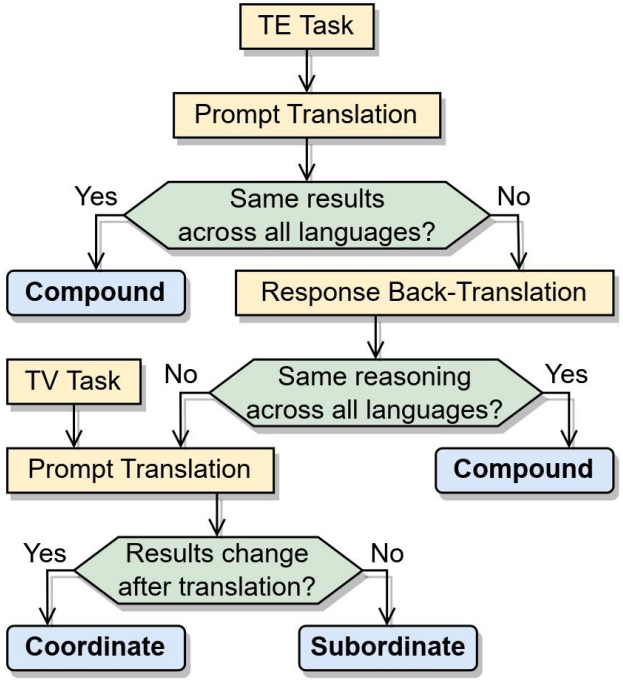
  Chinese Translation (no pun): 自行车不能独自站立，因为它有两个轮胎。

- Cover letter generation
  Example prompt:
  You are [name] from [school/institution] with [your Grade/GPA/Award]. You like [hobbies]. You want to join [company name]. Write a formal cover letter about: [restricted topics].

Question: What is the Multilingualism of LLMs?

- LLMs are known to be robust across languages on many tasks, but evaluations often focus on monolingual tasks.

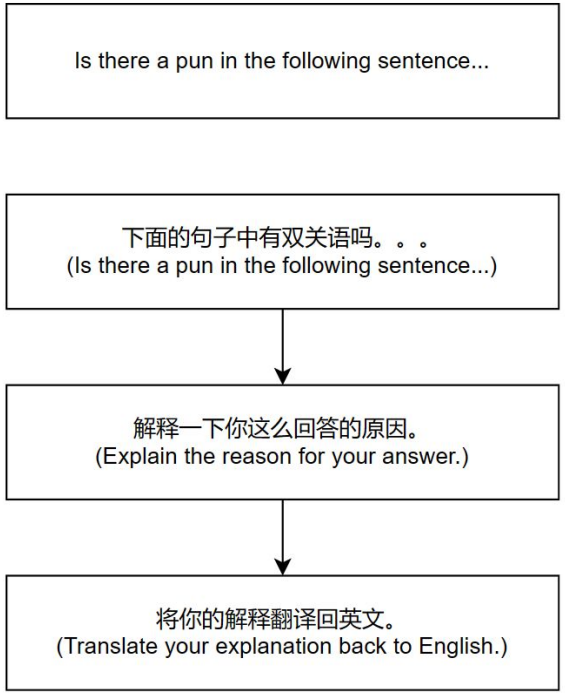
- We must understand the multilingual capabilities of LLMs in all languages.
  - How can we study the multilingual capabilities of LLMs?
  - What kind of mistakes can we expect LLMs to make?
  - How can we avoid the mistakes?



Approaches: PT and RBT

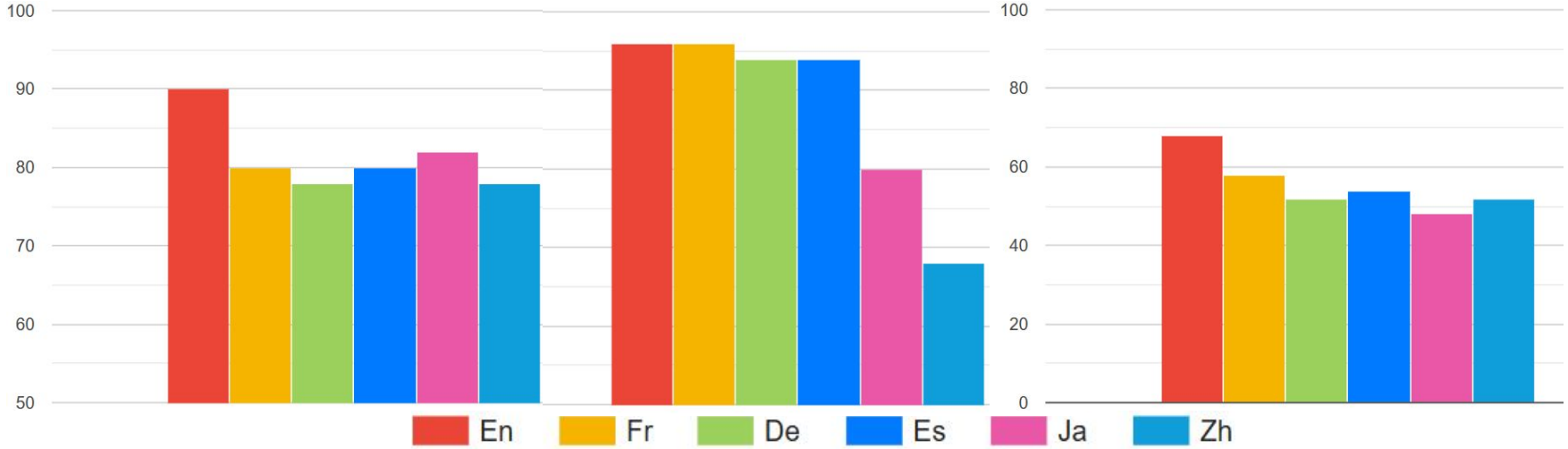
- Prompt Translation (PT):
  Translate monolingual datasets to generate parallel multilingual parallel data.

- Response Back-Translation (RBT):
  1. Translating prompts to the target language
  2. Prompting the LLM
  3. Prompt the LLM to generate an explanation
  4. Prompt for the translation of the explanation in the original language.



Results: Translation Equivariant Tasks

- **ChatGPT** is **better** at reasoning and retrieving knowledge given an English prompt, compared to prompting in other languages.

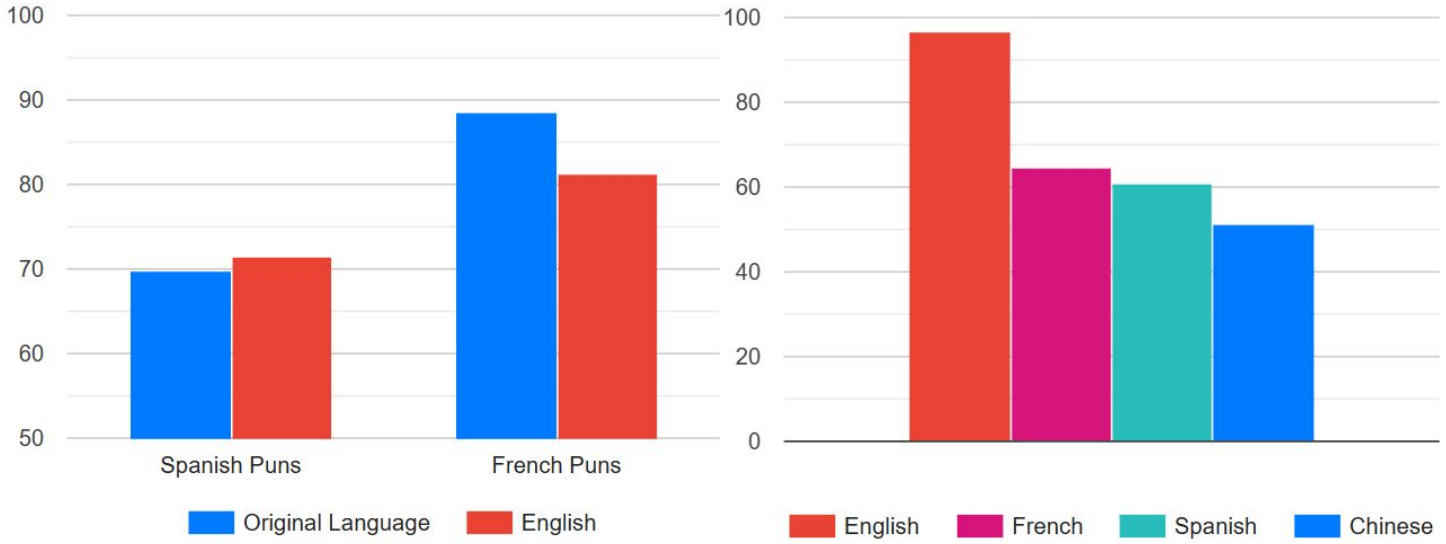


Accuracies in Math Reasoning, Knowledge Access, Common Sense Reasoning

We used gpt-3.5-turbo

Results: Translation Variant Tasks (1)

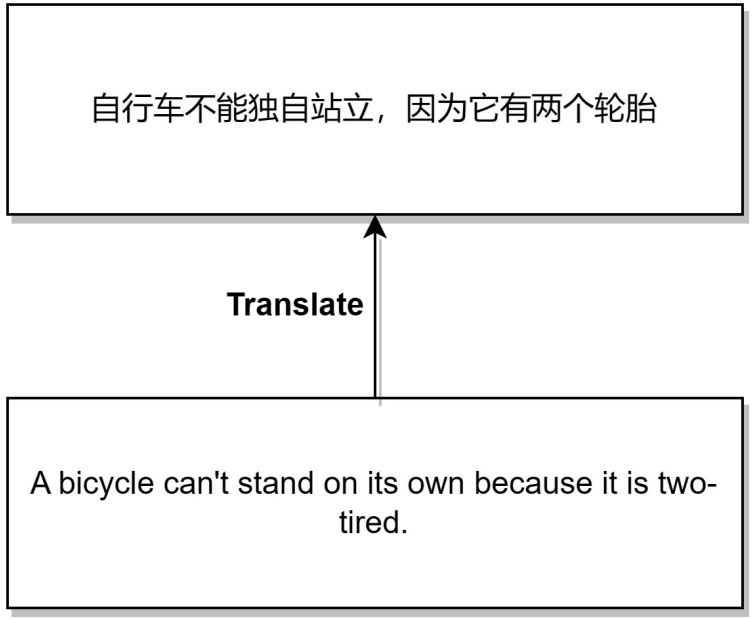
- Pun location accuracy for original and translated puns



Location Accuracies of Pun Detection

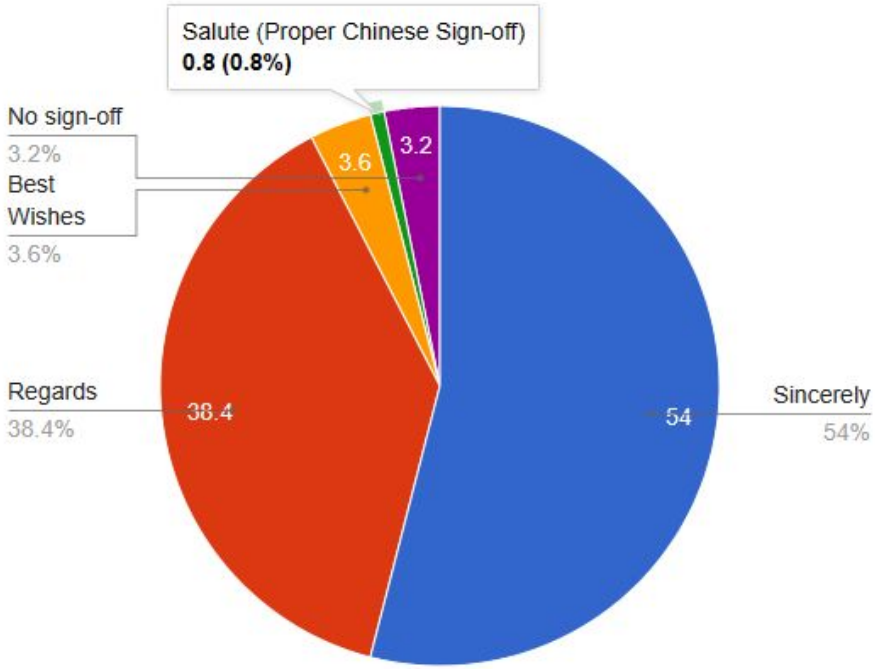
Case Study: Pun Detection

- The English sentence contains a pun: Two-tired -> Too tired.
- The Chinese translation of it does not contain any pun
- ChatGPT says there is a pun in the Chinese translation.
  Reason: the Chinese sentence has a pun because the English translation of it has a pun.



Results: Translation Variant Tasks (2)

- Task: cover letter generation.
- The correct letter sign-off (Salute) in Chinese is almost never used.
- ChatGPT's Chinese output reads like it was written in English and translated, providing strong evidence of subordinate multilingualism.



Conclusion

- We propose a novel method for studying the multilingualism of LLMs.
- ChatGPT exhibits behavior characteristic of subordinate multilingualism, with English as its "native language". This negatively impacts performance on non-English languages.
- Suggestions:
  - More parallel multilingual training data.
  - Incorporate multimodal (e.g. vision) training data.
  - Pursue the development of compound multilingual models.