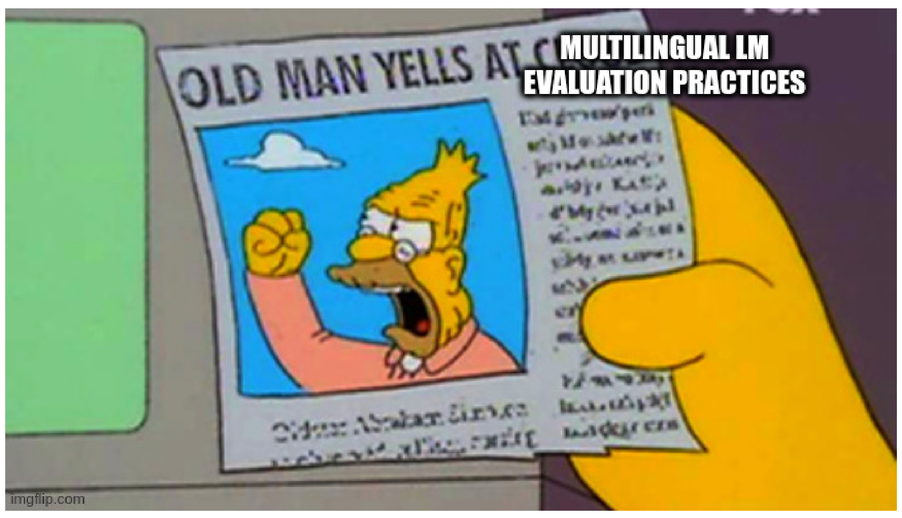
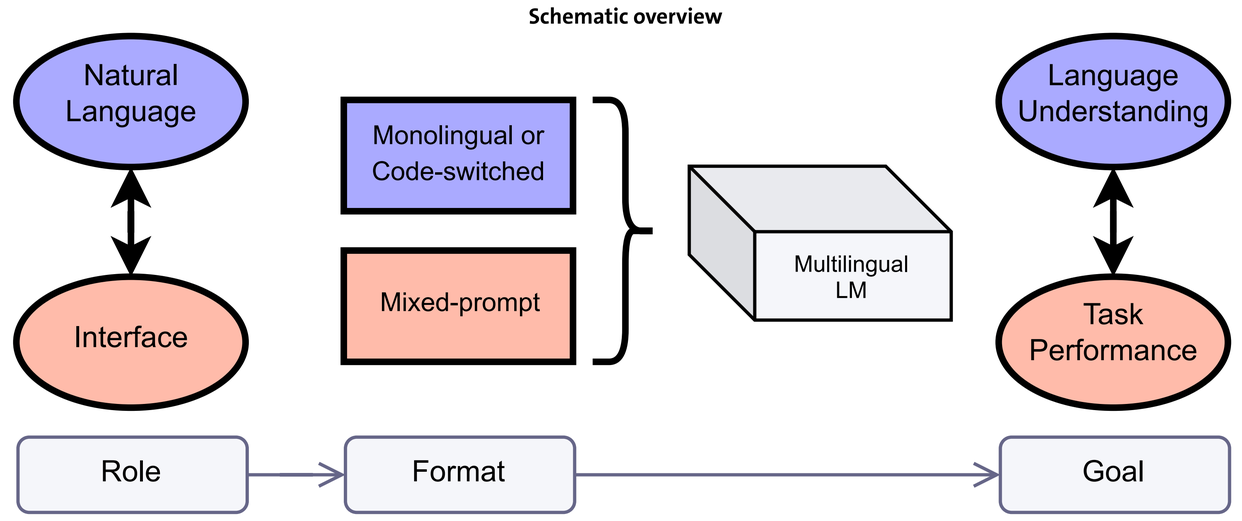
The Roles of English in Evaluating Multilingual Language Models

Wessel Poelman & Miryam de Lhoneux
LAGOM·NLP KU Leuven Belgium
wessel.poelman@kuleuven.be

The situation



Schematic overview



1. Background

- **Multilinguality** is gaining interest in NLP.
- **English** is predominantly used to prompt **multilingual** language models.
- Lack of **instruction tuning** data in other languages.
  → This results in two *roles* of English in evaluation.
  → Role as an interface and as a natural language.
  → Different goals: task performance versus language understanding.

2. Evaluation Setups

Task performance: a task as an end in itself.

- Working system with English $>$ poor performance or no system.
- Understanding English instructions and target language passages is not the same as *multilingual natural language understanding*.
- Auto-regressive LMs meant for direct interaction $\rightarrow$ usability issue.

$\rightarrow$ **Uses 'mixed-prompts' leading to superfluous evaluation issues!**

TL;DR: English is not a programming language.

Language understanding: a task as a proxy for understanding.

- Natural Language Understanding (NLU) benchmarks.
- Arguably includes the *instruction* domain.
- English by itself and in a *natural* mix.

$\rightarrow$ **Uses native prompts or natural code-switching: clear evaluation!**

3. Examples

The topic of the news **{sentence}** is **{topic}**

The topic of the news Bu oteller günün zenginlerinin ve ünlülerinin kalacağı yerlerdi ve çoğu zaman kaliteli yemeklere ve gece hayatına sahipti. is entertainment

1: Unnatural mixed-prompt of English and Turkish.

Taken from MaLa-500 evaluation setup (Lin et al., 2024).

You are a highly knowledgeable and intelligent artificial intelligence model answers multiple-choice questions about **{subject}**
Question: **{question}**
Choices:
A: **{choice1}**
B: **{choice2}**
C: **{choice3}**
D: **{choice4}**
Answer:

2: Superfluous effects of mixed-prompts cause imprecise evaluations:

- Script switching.
- Mixed-prompt switching *or* code-switching.
- English grammatical error correction.
- English instruction following.
- Answering high-school exam questions in the target language.

Taken from AfriMMLU evaluation setup (Adelani et al., 2024).

Translate this sentence from **{source}** to **{target}**
Source: **{source_sentence}**
Target:

Translate this sentence from German to Dutch
Source: Du gehst mir auf den Keks
Target:

Translate this sentence from Dutch to German
Source: tijd voor een bakje koffie
Target:

3: English is neither the source nor target language! It's an interface.

Taken from popular paper evaluating translation capabilities of OpenAI models (Hendy et al., 2023).

DE $\rightarrow$ NL (Dutch speaker)
Wat betekent "Du gehst mir auf den Keks" in het Nederlands?

NL $\rightarrow$ DE (Dutch speaker)
Hoe zeg je "tijd voor een bakje koffie" in het Duits?

$\rightarrow$ More natural code-switched prompt to achieve same result.

Acknowledgements

We thank the LAGoM-NLP group at KU Leuven for valuable discussions and paper recommendations. WP is funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096.

Paper

LinkedIn