The Roles of English in Evaluating Multilingual Language Models

Wessel Poelman and **Miryam de Lhoneux**
Department of Computer Science
KU Leuven, Belgium
{wessel.poelman, miryam.delhoneux}@kuleuven.be

Abstract

Multilingual natural language processing is getting increased attention, with numerous models, benchmarks, and methods being released for many languages. English is often used in multilingual evaluation to prompt language models (LMs), mainly to overcome the lack of instruction tuning data in other languages. In this position paper, we lay out two roles of English in multilingual LM evaluations: as an *interface* and as a *natural language*. We argue that these roles have different goals: *task performance* versus *language understanding*. This discrepancy is highlighted with examples from datasets and evaluation setups. Numerous works explicitly use English as an interface to boost task performance. We recommend to move away from this imprecise method and instead focus on furthering language understanding.

1 Introduction

With the increase of in-context, prompt-based evaluation of auto-regressive languages models (LMs, Brown et al., 2020), choices have to be made on how prompts are created. Specifically in multilingual evaluation, a crucial choice is in which language(s) prompts are written. In practice, English tends to be mixed with a target language with the explicit goal of increasing *task performance*. We argue this goal is different from furthering *language understanding*. In this position paper, we outline two roles of English at the core of this discrepancy and their implications.

Several works have highlighted methodological issues in multilingual evaluation setups (Artetxe et al., 2020; Ploeger et al., 2024). The dominance of English in natural language processing (NLP) has also been discussed repeatedly (Joshi et al., 2020; Ruder et al., 2022). With the increase of prompt-based evaluations of models, a new issue has appeared: English being used as an *interface*, rather than a *natural language*.

In recent work, Zhang et al. (2023) propose a taxonomy of prompt-based multilingual LM evaluations. They conclude that *"[the model] achieves higher performance when the task is presented in English."* This finding is consistent among a large number of papers (Shi et al., 2022; Huang et al., 2022; Fu et al., 2022; Lin et al., 2022; Asai et al., 2024; Etxaniz et al., 2024, inter alia). Resorting to using English like this is hardly surprising given that instruction tuning datasets are expensive to create and not readily available for most languages. Less surprising still is the finding that English performs well, as it is included in virtually all LMs. It does bring into question: what is being evaluated and what do we learn from this?

To illustrate: MaLa-500 (Lin et al., 2024) is a Llama 2-based model (Touvron et al., 2023) that underwent continued pre-training in over 500 languages. It is partially evaluated on a news topic classification task using SIB-200 (Adelani et al., 2024a), a dataset of $(sentence, topic)$ pairs in 205 languages. The model is prompted as follows:

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

Using the prompt with a Turkish[1] example gives:

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

This format is used across all 205 languages in few-shot setups from one to ten. This mixture of English and a target language is, arguably, not very 'natural'. We refer to this role of English as an *interface*, rather than a *natural language*. In the next sections, we outline these roles and why they are important to consider in multilingual evaluation.

[1] English translations of examples are in Appendix A.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 492–498
March 3-4, 2025 ©2025 University of Tartu Library

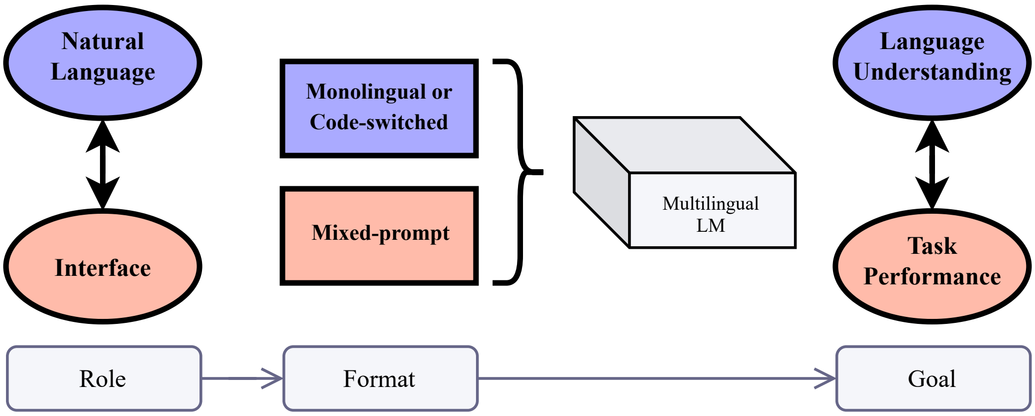

Figure 1 – Schematic overview of the different roles of English in multilingual LM evaluation.

2 Evaluation Goals

Language understanding. We take the common perspective that evaluation concerns a *task* which is used as a proxy for *understanding*. This is exemplified by the *natural language understanding* (NLU) label many datasets and models adhere to (including SIB-200). A news topic classification task shows that the model (arguably) 'understands' some of the differences between news categories. A model that rewrites, translates or summarizes 'understands' both task instructions and target passages. In a multilingual setting, the understanding of interest is *generalizability* across languages; a model performing a task in a target language supposedly *understands* something about that language. This is then applied to multiple languages. We refer to this as 'multilingual natural language understanding' (MLU). Specifically, we use MLU to mean 'understanding a target language is part of multilingual natural language understanding.'[2]

Understanding English by itself and understanding a *natural* mix of English and another language are both part of MLU. The latter enters the domain of code-switching: the phenomenon where a speaker fluently switches between multiple different languages during the same conversational turn (Milroy and Muysken, 1995).[3]

The MaLa-500 prompt mixes English and a target language. However, it is hard to classify this as code-switching, as the switch is hardly natural, especially in a few-shot setup. Rather than a *natural language* that tells something about *language understanding*, English is used as an *interface* to the LM with the goal of increasing *task performance*. We refer to this mixing as a *mixed-prompt*.

Task performance. Another widespread perspective on evaluation in (multilingual) NLP considers performance on a task as an end in itself.[4] If we want to classify news topics in a practical application operating in a multilingual setting, what a model supposedly understands or how well it models a particular language is of little value. What matters is the system performing its task adequately across languages. Without using English, the system might not even work at all. This is a common justification; mixing in English is arguably better than not having a system at all.

While practical, this perspective is seemingly at odds with the many tasks and datasets that present themselves under the aforementioned label of language *understanding*. Additionally, task performance as the sole goal introduces a usability issue. Auto-regressive LMs are increasingly meant to be directly interacted with (a *natural* language interface). If we have to resort to a mixed-prompt for the system to even function, it means the user has to be able to write English and get familiar with this unnatural mixing of languages.

Figure 1 summarizes our argument and terminology. Next, we provide more details regarding the discrepancies between using English as an interface versus using it as a natural language.

[2] We are aware this (ab)use of terminology is not standard.

[3] Some differentiate between code-switching and code-mixing, we do not make a distinction. For an overview of code-switching in NLP, we refer to Winata et al. (2023).

[4] We thank two reviewers for suggesting to put more emphasis on this perspective.

3 Evaluation Methods

As mentioned in §1, a large body of contemporary research in multilingual NLP focuses on prompting methods. Common evaluation setups range from (i) prompts fully in a target language, to (ii) English instructions with task-specific passages in the target language, to (iii) translating all text into English before presenting it to a model.[5] None of these works refer to this mixture as being code-switched text. All conclude that a mixture of English and a target language (a mixed-prompt) generally results in the best task performance. In this section we show why a mixed-prompt is an inherently imprecise method to use in evaluation, even if maximizing task performance is the goal.

If we use a prompt fully in a target language, we are clearly evaluating part of MLU. A mixed-prompt introduces *additional factors* that are evaluated that are neither the task nor MLU. We illustrate this from two angles: the representation of the prompt and fortuitous issues from unnaturally mixing English and a target language.

Consider how to evaluate a multilingual masked language model on the news classification task. A classification layer is added to a pre-trained model to predict the topic labels; it sees label *indices* that are consistent across languages. The labels are language-agnostic for the model (i.e., detached from natural language). The evaluation method and goal are clear: mapping a target language sequence to one of these indices. There are no additional signals influencing this process.

In a prompting setup, the representation of the labels can either be language-agnostic (numbers, letters, symbols, etc.), or not (English words, target language words, etc.). These options result in any number of *tokens*, which will have different representations within the model, unless specifically accounted for. In many multilingual evaluation prompts, the classification labels are English words (such as in the MaLa-500 example). Without target language words or (to an extent) language-agnostic labels, the evaluation method and goal will be inherently imprecise.

In addition to the different representation, more than just the task is evaluated with a mixed-prompt setup. To illustrate this, consider the following setup from the AfriMMLU subtask of IrokoBench (Adelani et al., 2024b):

> 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:

The prompt and `subject` are always in English, the `question` and `choices` in the target language. With this setup, more is tested than just a task in a target language:

- Code-switching, if this is considered natural, or unnatural 'mixed-prompt' switching.
- Script-switching, if the target language uses a non-Latin script (which applies to Amharic in IrokoBench, using the Ge'ez script).
- Instruction following in English.
- Grammatical error correction in English.[6]
- Answering high-school level exam questions in the target language.

With these mixed-prompts, we arguably do not test MLU, as that would entail a native target language prompt. At the same time, we test more than just the task, even though that is the explicit goal of using English in this way.

While we only discussed classification tasks until now, our argument also applies to other types of tasks. Consider the following zero-shot machine translation prompt from Hendy et al. (2023):

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

The prompt is always in English, the `source` and `target` are English words referring to the languages, and the `source_sentence` is in the target language. Filled in, it looks like this:

> # DE → NL
> Translate this sentence from German to Dutch
> Source: Du gehst mir auf den Keks
> Target:
>
> # NL → DE
> Translate this sentence from Dutch to German
> Source: tijd voor een bakje koffie
> Target:

[5] We do not further discuss 'translate everything' as this resembles evaluating English as a *natural language*.

[6] We have notified the AfriMMLU authors about this. The typo is in the prompt in the paper and in the *lm-evaluation-harness* (Biderman et al., 2024), which is used to obtain their results: https://github.com/EleutherAI/lm-evaluation-harness/blob/7882043b4ee1ef9577b829809c2f4970b0bdba91/lm_eval/tasks/afrimmlu/direct/utils.py.

The authors mention they *"explore prompt selection strategies along two dimensions: quality and relevance"*, but do not mention target language prompts. To underline the *interface* role of English: it is neither the translation source nor target here. Hendy et al. (2023) mention that *"keeping the prompt format the same allows us to potentially leverage the benefits of the underlying instruction finetuning protocol to the full extent."* This makes explicit the goal of *task performance*. Prompting a model to translate a sentence is easily done in a manner that more closely aligns with the goal of MLU, does not use English, and is closer to natural code-switching:

> # DE → NL (Dutch speaker)
> Wat betekent "Du gehst mir auf den Keks" in het Nederlands?
>
> # NL → DE (Dutch speaker)
> Hoe zeg je "tijd voor een bakje koffie" in het Duits?

4 Why does this matter?

Interacting with computers in a natural manner is arguably the ultimate goal of numerous subfields of computer science. Work on natural language interfaces to information systems dates back decades (Winograd, 1972; Waltz, 1978). LMs bring us ever closer to this goal. However, in a multilingual setting, it is important to consider what *natural language* is, what is being evaluated, and what promises are sold. Next, we outline the implications of the *interface* versus *natural language* roles on evaluation practices.

Interface. Let us start with the role in which English is akin to a programming language.[7] We need an interface to communicate with a system, in a way the system can understand. We have seen that mixed-prompts are used to get the system to perform better on a given task. Given the scarcity of instruction tuning datasets and the costs involved in creating these, it is understandable that this is a common (albeit sometimes implicit) perspective. English becomes the 'programming' language that glues target language passages together and makes the system perform a task. Programming languages also predominantly use English labels for their keywords. However, if the keyword for a `while` loop happens to be `mientras` or `kjsdfk` is irrelevant for its function. These are natural language-agnostic as the meaning (as interpreted by a compiler or interpreter) does not change. Variable names and keywords can be chosen arbitrarily.[8] This is not the case with prompting, which is sensitive to slight changes, both in English (Sclar et al., 2023) and multilingual setups (Zhang et al., 2023; Asai et al., 2024).

Additionally, evaluation setups that use English as an interface introduce knowledge leakage from English to the target language. This is, again, with the explicit goal of improving task performance.[9] Being able to understand English instructions is not the same as being able to understand target language instructions. If English truly was a programming language, this would not matter, as the meaning of the instructions would be separate from the meaning of the target language passages. Given that English is a natural language, this *de facto* means more is evaluated than just the task. Consequently, such evaluations are imprecise at best, as shown in §3.

Prompt-based evaluations should extend MLU to the *instruction* domain. A mixed-prompt setup claiming to test *"multilingual understanding"* might more accurately be described as *"understanding English instructions interleaved with passages from target language(s), albeit not in a natural code-switching setup."*

Natural language. When we consider the other role of English in multilingual prompt-based evaluation, we should treat it the same as any other language. The 'Multilingual Exemplars' setup from Shi et al. (2022) is a creative interpretation of this perspective. In this few-shot setup, the model sees various examples, all in *different* languages. The final question is asked in the target language. A setup like this extends the definition of 'multilingual language understanding' to the extreme. It becomes harder to interpret what a multilingual model knows about any individual language in this context, but English is certainly not an interface, it is a natural language like all others.

A less extreme setup would simply use native, target language prompts or natural code-switched prompts. This is costly, but it aligns much bet-

[7] Also reflected in this famous post: https://x.com/karpathy/status/1617979122625712128

[8] Within the specifications of the programming language.

[9] Knowledge leakage also explicitly happens in parameter sharing (Zeman and Resnik, 2008) or cross-lingual transfer (Philippy et al., 2023). However, these methods are fundamentally different from mixed-prompts as they (i) treat English as a natural language, and (ii) target knowledge sharing at the training or finetuning phase, not the evaluation phase.

ter with the goal of multilingual natural language understanding. Indeed, several works specifically explore this direction (Köpf et al., 2023; Singh et al., 2024). This approach clearly tests multilingual language understanding, including the instruction domain. If performance on a particular task in a particular language is lagging behind, or not working at all, it means focus should be put on addressing the core of these issues (e.g., data or modeling). Ideally, we should not resort to imprecise methods to boost task performance.

5 Conclusion

In this position paper we outline two roles of English in multilingual language model evaluation: as an *interface*, with the goal of *task performance*, and as a *natural language*, with the goal of *language understanding*. We (i) list works that incorporate English with the explicit goal of boosting task performance, even in tasks such as translation where it is neither the source nor target, underlining the *interface* role, (ii) show that mixing English with a target language in a *mixed-prompt* is unnatural (i.e., not code-switching), and (iii) outline why the interface role is an imprecise choice when evaluating multilingual language understanding of language models.

Additionally, we argue that using a mixed-prompt tests *more* than just performance on a certain task. Because English is a natural language and not a programming language, using it in a mixed prompt will inherently lead to fortuitous factors such as (un)natural switching between languages or scripts, grammatical error correction, and more. This all results in imprecise or misleading evaluations, even if the ultimate goal was to evaluate and improve task performance.

We finally contrast the implications of the two roles on evaluation practices. We recommend to move away from using English as an interface in multilingual evaluations and ultimately advocate for the goal of *language understanding*.

Acknowledgments

WP is funded by a KU Leuven Bijzonder Onderzoeksfonds C1 project with reference C14/23/096. We thank the LAGoM-NLP group at KU Leuven for valuable paper recommendations and Mahdi Dhaini for reviewing an early draft of this paper. We also thank the reviewers for their constructive comments.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024a. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. 2024b. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. *arXiv preprint, arXiv:2406.03368v1*.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A Call for More Rigor in Unsupervised Cross-lingual Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the Trenches on Reproducible Evaluation of Language Models. *arXiv preprint, arXiv.2405.14782v2*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin

Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do Multilingual Language Models Think Better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot Prompt: Multilingual Multitask Prompt Training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9919–9935.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint, arXiv:2302.09210v1*.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot Cross-lingual Transfer of Prompt-based Tuning with a Unified Multilingual Prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul Es, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. OpenAssistant Conversations - Democratizing Large Language Model Alignment. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. MaLA-500: Massive Language Adaptation of Large Language Models. *arXiv preprint, arXiv:2401.13303v2*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Lesley Milroy and Pieter Muysken, editors. 1995. *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching*. Cambridge University Press.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is "Typological Diversity" in NLP? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint, arXiv:2307.09288v2*.

David L. Waltz. 1978. An English language question answering system for a large relational database. *Communications of the ACM*, 21(7):526–539.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Daniel Zeman and Philip Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927.

A Examples

The examples containing Turkish, Dutch or German are repeated here with English translations.

SIB-200 (sample 755):

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

The topic of the news *These hotels were where the rich and the famous of the day would stay, and often had fine dining and nightlife.* is entertainment

Interface translation examples:

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

# DE → NL
Translate this sentence from German to Dutch
Source: *You're getting on my nerves*
Target:

# NL → DE
Translate this sentence from Dutch to German
Source: tijd voor een bakje koffie
Target:

# NL → DE
Translate this sentence from Dutch to German
Source: *time for a cup of coffee*
Target:

Natural translation examples:

# DE → NL (Dutch speaker)
Wat betekent "Du gehst mir auf den Keks" in het Nederlands?

# DE → NL (Dutch speaker)
What does "Du gehst mir auf den Keks" *mean in Dutch?*

# NL → DE (Dutch speaker)
Hoe zeg je "tijd voor een bakje koffie" in het Duits?

# NL → DE (Dutch speaker)
How would one say "tijd voor een bakje koffie" *in German?*