Thank You, Stingray: Multilingual Large Language Models Can Not (Yet) Disambiguate Cross-Lingual Word Sense

Anonymous ACL submission

Abstract

Multilingual large language models (LLMs) have gained prominence, but concerns arise regarding their reliability beyond English. This study addresses the gap in cross-lingual semantic evaluation by introducing a novel benchmark for cross-lingual sense disambiguation, StingrayBench[1]. In this paper, we demonstrate using false friends—words that are orthographically similar but have completely different meanings in two languages— as a possible approach to pinpoint the limitation of cross-lingual sense disambiguation in LLMs. We collect false friends in four language pairs, namely Indonesian-Malay, Indonesian-Tagalog, Chinese-Japanese, and English-German; and challenge LLMs to distinguish the use of them in context. In our analysis of various models, we observe they tend to be biased toward higher-resource languages. We also propose new metrics for quantifying the cross-lingual sense bias and comprehension based on our benchmark. Our work contributes to developing more diverse and inclusive language modeling, promoting fairer access for the wider multilingual community.

1 Introduction

Multilingual large language models (LLMs) have become integral tools in a variety of tasks and languages (Bang et al., 2023; Yong et al., 2023; Zhang et al., 2023a; Lovenia et al., 2024; Cahyawijaya, 2024; Cahyawijaya et al., 2024). While these LLMs have remarkable capabilities, there are growing concerns about the reliability of their responses especially in languages other than English. Most evaluations address cross-lingual generalization in LLMs by assessing their ability on the set of downstream tasks as the one used in English (Cahyawijaya et al., 2021; Adelani et al., 2023; Kabra et al., 2023; Zhang et al., 2023b; Adelani et al., 2024; Cahyawijaya et al., 2024; Zhang and Eickhoff, 2024), many even directly translated from the source corpora (Hu et al., 2020; Cahyawijaya et al., 2021; Winata et al., 2023; Cahyawijaya et al., 2023a; Bandarkar et al., 2024; Singh et al., 2024). These evaluations reflect the cross-lingual generalization in the downstream application level, but fail to capture the basic understanding of semantic meaning across different languages. This lack of semantic understanding further extends to the unexplained bias of multilingual LLMs towards certain languages or language families which causes the LLMs to respond in their preferred languages, leading to a significant misrepresentation of users' intent (Nomoto, 2023; Nomoto et al., 2024).

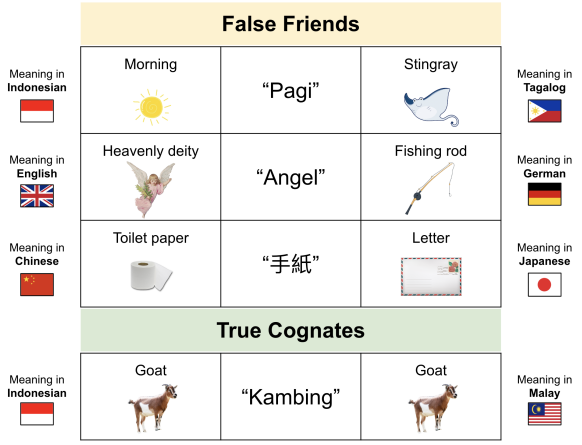


Figure 1: Our work explores two linguistic phenomena known as **false friend** and **true cognate**, and highlights the limitation of LLMs on understanding cognate indicating the pitfall on cross-lingual disambiguation.

Our work aims to explore the cross-lingual evaluation of semantic meaning in LLMs and understand its underlying causes. We focus on the concept of "false friends", which are words or phrases that sound similar in two languages but have distinct meanings[2] and "true cognates", which are

[1]To ensure reproducibility, we will release our benchmark and evaluation suite under the CC-BY-SA 4.0 license.

[2]Our title "Thank You, Stingray" is a playful reference

words or phrases that sound similar in two languages and share the same meaning[3]. We create data instances containing false friends and true cognates as described in Figure 1. Using these concepts, we construct the first benchmark for measuring cross-lingual semantic understanding in LLMs dubbed as StingrayBench. By analyzing LLMs performances on StingrayBench containing multiple language pairs, we assess whether they exhibit language-selection bias through the task of cross-lingual sense disambiguation with new metrics and present future research directions to mitigate these biases. Our contributions and the significance of this work can be summarized as follows:

- We propose **StingrayBench**, the first benchmark for measuring the cross-lingual sense disambiguation in LLMs covering four distinct language pairs, i.e., Indonesian-Malay (ID-MS), Indonesian-Tagalog (ID-TL), Chinese-Japanese (ZH-JA), and English-German (EN-DE).
- We introduce a method to measure cross-lingual sense comprehension and bias in LLMs by introducing **stringray plot** and two evaluation metrics for measuring cross-lingual sense understanding, i.e., **cognate bias** and **cognate comprehension score**.
- We showcase the generalization of the cognate bias phenomena to multiple multilingual LLMs in diverse language pairs, demonstrating its broader impact and severity in existing multilingual LLMs.

2 Related Works

2.1 Cognates and False Friends in NLP

Homologous words that show systematic sound correspondences indicating common ancestry are known as cognates (Atkinson, 2013). For example, *baru* in Malay and *bago* in Tagalog are cognates based on the systematic *r-g* sound correspondence, both meaning 'new'. However, cognates do not necessarily have the same meaning, as is the case with *bibir* meaning 'lip' in Malay and *bibig* 'mouth' in Tagalog, which show the same *r-g* correspondence. The study of cognacy contributes to understanding the historical lineage of languages and the reconstruction of proto-languages (Campbell, 2013).

Many recent works focus on the identification of cognates in genetically related languages (Batsuren et al., 2019, 2021; Bafna et al., 2022; Dinu et al., 2023; Akavarapu and Bhattacharya, 2024). One of the factors that make cognate identification non-trivial is the presence of false friends (or false cognates). False friends are words that are orthographically or phonetically similar but do not share the same meaning (Allan, 2009). While many false friends are indeed cognates, some are not true cognates and can be mistaken for cognates. For example, an Indonesian-Malay false friend *polisi* 'police (Indonesian), policy (Malay)' traces back to different ancestor languages, i.e. Dutch (*politie* 'police') and English (*policy*). Besides posing a challenge to cognate identification, false friends also constitute a major obstacle for translators, language learners and especially machine translation systems. Studying false friends is not easy because it requires bilingual proficiency and as a result, false friends have received little attention. Current studies on false friends focus on their collection and identification Ljubesic and Fišer (2013); Castro et al. (2018); Uban and Dinu (2020). In our paper, we deal with false friends in multiple language pairs and use them as a tool to understand the proficiency of multilingual large language models.

2.2 Word Sense Disambiguation

Word Sense Disambiguation (WSD) task aims to determine the correct meaning of a polysemous word in a given context (Bevilacqua et al., 2021, 2020; Blevins and Zettlemoyer, 2020) The task has also been extended to a multilingual setting (Navigli et al., 2013; Pasini, 2021; Pasini et al., 2021; Su et al., 2022), leveraging multilingual lexical knowledge bases (Navigli and Ponzetto, 2012; Bond and Foster, 2013). It has been a challenging task in NLP since its early recognition by Weaver (1949) and remains critical in recent works that investigate "the curse of multilinguality" (Conneau, 2019; Berend, 2023) and universal representations across languages (Wu et al., 2019; Wendler et al., 2024; Ferrando and Costa-jussà, 2024; Zhang et al., 2024). Built upon previous studies, our work explores the ability of multilingual LLMs to disambiguate word sense across languages. The closest related task is the Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2018; Raganato et al., 2020), where they ask to classify the word usage given two distinct contexts in the same language. In

to a false friend phrase "Selamat Pagi", which means "Good morning" in Indonesian but means "Thank you, stingray" in Tagalog, which might bring confusion to multilingual LLMs.

[3]We use the term **common words** as the umbrella term to refer to both true cognates and false friends.

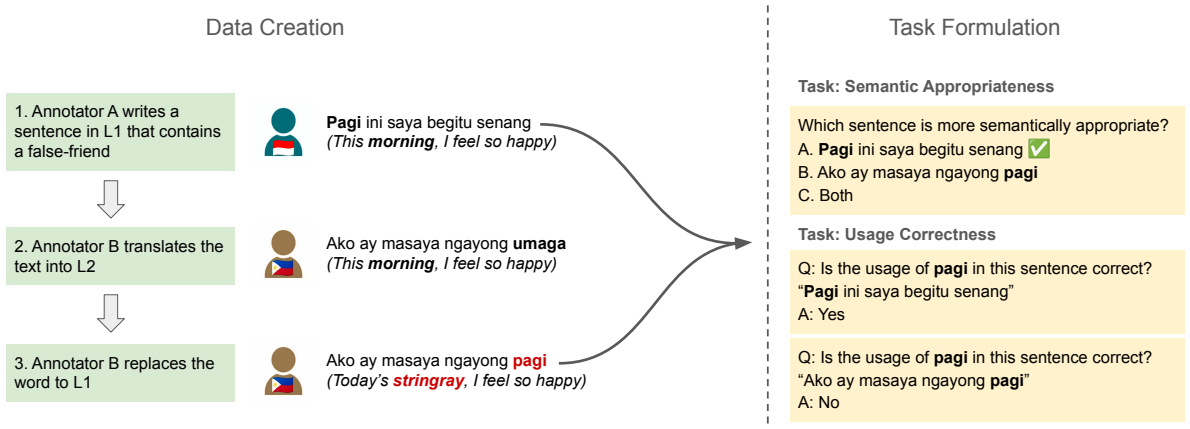


Figure 2: Annotation and data formulation pipeline of StingrayBench. Our annotation consists of a 3-step process that requires two annotators, one for each language of the language pair. In addition, we provide the English translation of the correct sentence for better accessibility to StingrayBench.

contrast, we embed the word in parallel contexts across different languages, with only one sentence being semantically correct. By evaluating whether LLMs align the word's meaning with the appropriate language-specific context, we can assess their multilingual capabilities and detect potential language selection bias.

3 StingrayBench

3.1 Dataset Construction

To construct StingrayBench, native speakers of Chinese, English, German, Indonesian, Japanese, Malay, and Tagalog are asked to list down common words that existed between the following language pairs together with their meanings: English-German, Indonesian-Malay, Indonesian-Tagalog, and Chinese-Japanese. For this, the following resources dealing with false friends are consulted: Wiktionary's lists of false friends,[4] *Kamus Komunikatif Nusantara: Indonesia-Malaysia, Malaysia-Indonesia* (Mohd Sharifudin bin Yusop and Al Mudra, 2015), and *Kamus Kata: Bahasa Melayu Malaysia-Bahasa Indonesia* (Rusdi Abdullah, 2016). Some words in these resources are rejected as they turned out not to be false friends after scrutiny. Moreover, it is not always easy to find common words with identical spellings and characters, except for Indonesian-Malay. Therefore, we allow the use of words differing in capitalization (e.g. *arm-Arm*) in English-German, the use of words with one edit distance (e.g. *aku-ako*) in Indonesian-Tagalog, and the use of words with different characters developed from the same origin (e.g. 图书馆-図書館) in Chinese-Japanese. These common words are then segregated into false friends (same word with different meanings) and true cognates (same word with the same meaning).

For each word, annotators would construct a sentence that uses that word in their native language. An English translation would then be written by the annotator. The annotator of the other language in the pair would then construct a sentence in their native language that follows the meaning of the sentence in the first language and/or that of the English translation. For sentences involving a false friend, an accurate translation would not employ the target false friend word but a different word that expresses the intended meaning in the language. Hence, an additional step of replacing the latter word with the target false friend word is required, which produces semantically odd sentences.

For example, in Indonesian-Tagalog, for the word *pagi* meaning 'morning' in Indonesian and 'stingray' in Tagalog. The Indonesian annotator would construct *Pagi ini saya begitu senang* as the sentence. The English translation would be *This morning, I feel so happy*. The Tagalog annotator would then translate it as *Ako ay masaya ngayong* ***umaga*** and replace the word meaning 'morning', i.e., *umaga*, by the target false friend word *pagi* to produce *Ako ay masaya ngayong* ***pagi***, which means 'Today's stingray, I feel happy' and is semantically odd in Tagalog.

In most cases, each false friend will have two entries in the final dataset corresponding to one correct and one incorrect usage of that word. However, in some cases, a false friend has only one

[4] https://en.wiktionary.org/wiki/Category:False_cognates_and_false_friends

Subset	#True Cognate	#False Friend	#Total
EN-DE	98	98	196
ID-MS	52	134	186
ID-TL	58	100	158
ZH-JA	51	114	165
Total	**259**	**446**	**705**

Table 1: Statistics of our StingrayBench.

entry. This happens for partial cognates: when the word shares the same meaning in two languages but has an additional meaning in one language but is absent in the other. For example, *pelatih* means 'trainer' in both Indonesian and Malay, but it has another meaning in Malay, but not in Indonesian, i.e. 'trainee'. Each true cognate will only have one entry as both native language sentences translate to each other correctly. The detailed annotation guideline is provided in Appendix A. The statistics of the StingrayBench are described in Table 1.

3.2 Task Formulation

Using the sentences collected above, we propose two task formulations of different semantic granularities as follows. Notice that we prompt the model in English as it is language-neutral for most of the language pairs except for the English-German case.

Semantic Appropriateness Given the data construction as described above, we want to test the models' competence for sentence comprehension. In this task, we prompt the model with: *Which sentence is more semantically appropriate?*. The first two options are the two sentences for the language pair respectively. A third option, that both sentences are appropriate, is also included. It is the correct option for the true cognates scenario but also serves as a confounding option for the false cognates subset. We provide the example of the prompt and the target completion for the semantic appropriateness task in Figure 2.

Usage Correction In this task formulation, we emphasize the usage of the specific cognate words by prompting the models with: *Is the usage of [WORD] in this sentence correct? [SENTENCE]*. We expect this task to be simpler as 1) the options are binary with no confounding options; and 2) specific cognates are mentioned in the prompt which potentially serves as a task hint. We provide the example of the prompt and the target completion for the usage correction task in Figure 2.

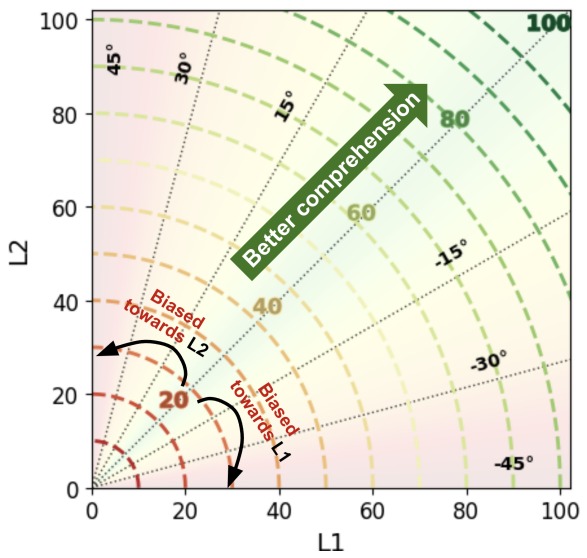


Figure 3: Stingray plot is a 2D scatter plot where the X-axis and Y-axis represent the model performance on StingrayBench across each language. The **cognate bias score** towards a particular language is measured based on the angular distance of the data point (e.g. the model is unbiased if it has equally good performance for either language). The **cognate comprehension score** is measured based on the point's magnitude.

3.3 Measure of Cognate Understanding Ability in LLMs

We define cognate understanding as the ability of LLMs to be able to correctly comprehend the semantic meaning of a cognate for both "true cognate" and "false friend". This is done through probing LLMs with questions that ensure the understanding of LLMs to the semantic nuances of the cognate, which can be either "true cognate" or "false friend", in the context of the relevant language pairs as described in Section 3.2. Using the tasks in StingrayBench, we measure the per-language accuracy of LLMs on each task and conduct further analysis as described below.

Stingray Plot To measure cognates understanding ability of LLMs on a certain language pair $< L_1, L_2 >$, we need to take into account the cognate understanding quality on both L_1 and L_2. To do so, we derive our analysis based on a 2-dimensional vector space and introduce the Stingray plot. As shown in Figure 3, the Stingray plot presents two different contours: (1) a U-shaped angular contour with a minimum value of 0 at either 0° and 90° angle and a maximum value of 1 at 45° angle; and (2) the radial contour with a minimum value of 0 at the bottom left corner and a maximum value of 100 at the top right corner. Using this

Model Name	Model Size	Supported Lang.
Monolingual / Bilingual LLMs		
ChatGLM2	6B	en, zh
Yi-1.5	9B, 34B	en, zh
Phi-3	3.8B, 7B, 14B	en
Cendol LLaMA-2	7B	id, (en, de)*
Cendol mT5	3.7B	id, (en, de, zh, ja, ms, tl)*
Multilingual LLMs		
SeaLLM v3	7B	en, zh, id, ms, tl
SEA-LION v2.1	8B	en, zh, id, ms, tl
BLOOMZ	0.6B, 1.1B, 1.7B, 3B, 7B	en, zh, id
mT0	0.3B, 0.6B, 1.2B, 3.7B, 13B	en, de, zh, ja, id, ms, tl
Aya-101	13B	en, de, zh, ja, id, ms, tl
Aya-23	8B, 35B	en, de, zh, ja, id, ms, tl
QWEN-2.5	0.5B, 1.5B, 3B, 14B, 32B	en, de, zh, ja
Command-R	35B	en, de, zh, ja, (id)*
GPT-4o Mini	-	en, de, zh, ja, (id, ms, tl)*
Llama-3.1	8B, 70B	en, de
Llama-3.2	1B, 3B	en, de

Table 2: List of LLMs incorporated in our experiment. For language codes, we adopt the ISO 639-3 standard. Asterisk (*) denotes that the language is not officially supported or is only included in the pre-training phase.

characteristic of the Stingray plot, we develop two metrics for measuring cognate understanding, i.e., **cognate bias** and **cognate comprehension**.

Cognate Bias Score Given a language pair, an LLM can perform well in identifying cognates in one, but poor in the other. In this case, we can expect that the model has a certain degree of understanding bias in one language. An unbiased LLM should yield similar performance on both languages, while an extremely biased LLM should perform well on one, and close to random estimator for the other. To quantify the **cognate bias score**, we follow the U-shaped angular contour in the Stingray plot. Specifically, we measure the angular distance between the $< L_1, L_2 >$ performance of an LLM with the 45° angle. To disambiguate between bias to L_1 and L_2, we incorporate the sign such that a negative distance indicates a bias towards L_1, while a positive distance indicates a bias towards L_2. Lastly, we normalize the range of the **cognate bias score** by linearly scaling from the original range of $[-\frac{\pi}{4} \ldots \frac{\pi}{4}]$ to $[-1.0 \ldots 1.0]$.

Cognate Comprehension Score Cognate bias shows the understanding of one LLM in a certain language, but it does not reflect the proficiency of LLM in understanding cognates. For instance, when an LLM behaves like a random estimator in both languages, it will yield similar accuracy scores (50% for binary classification, 33% for ternary classification, etc) in both languages. In this case, the LLM does not seem to exhibit much cognate bias, but it does not imply that the LLM has an exceptional cognate understanding. To quantify the cognate understanding ability, we introduce the **cognate comprehension score**. A perfect cognate comprehension score indicates that the LLM is unbiased and performs well in both languages. The cognate comprehension is implemented by simply calculating the magnitude of the $< L_1, L_2 >$ vector and normalizing the range into $[0 \ldots 1]$ by dividing the magnitude with $\sqrt{2}$. Note that, when the LLM yields a perfect score on L_1 and 0 on L_2, the performance only achieves $\sim$70.71% cognate comprehension score, further improvement from this point will also reduce the bias of the LLM.

4 Experiment Setting

4.1 Data Subsets

We utilize the collected StingrayBench for our evaluation which covers four language pairs, i.e., English-German, Indonesian-Malay, Indonesian-Tagalog, and Chinese-Japanese. For each language, we split the data into two different subsets based on the phenomenon observed, i.e., true cognate and false friend subsets. As there are only limited amount of data, we aggregate the score from multiple tasks to improve the reliability of the LLMs prediction. The statistics of the StingrayBench per language pair and per subset are shown in Table 1.

4.2 Model

Our evaluation covers a wide variety of LLMs, from monolingual, bilingual, and multilingual LLMs. For multilingual LLMs, we incorporate BLOOMZ (Le Scao et al., 2023; Muennighoff et al., 2023), mT0 (Muennighoff et al., 2023), Aya-101 (Singh et al., 2024; Üstün et al., 2024), Aya-23 (Aryabumi et al., 2024), Qwen-2.5 (Yang et al., 2024; Team, 2024), Command-R (Cohere For AI, 2024a,b), and GPT-4o mini (OpenAI et al., 2024). We also explore LLMs with lower language coverage or specifically adopted for certain languages including Phi-3 (Abdin et al., 2024), Cendol LLaMA-2 (Touvron et al., 2023; Cahyawijaya et al., 2024), Cendol mT5 (Xue et al., 2020; Cahyawijaya et al., 2024), SEALLM v3 (Nguyen et al., 2024), SEA-Lion v2.1 (Ong and Limkonchotiwat, 2023), ChatGLM2 (GLM et al., 2024), and Yi (AI et al., 2024). We exhaustively explore different size variations of each LLM with the scale ranging from 0.3B to 70B parameters to better understand the effect of scaling on the cognate understanding of LLMs. The list of LLMs covered in our study is shown in Table 2.

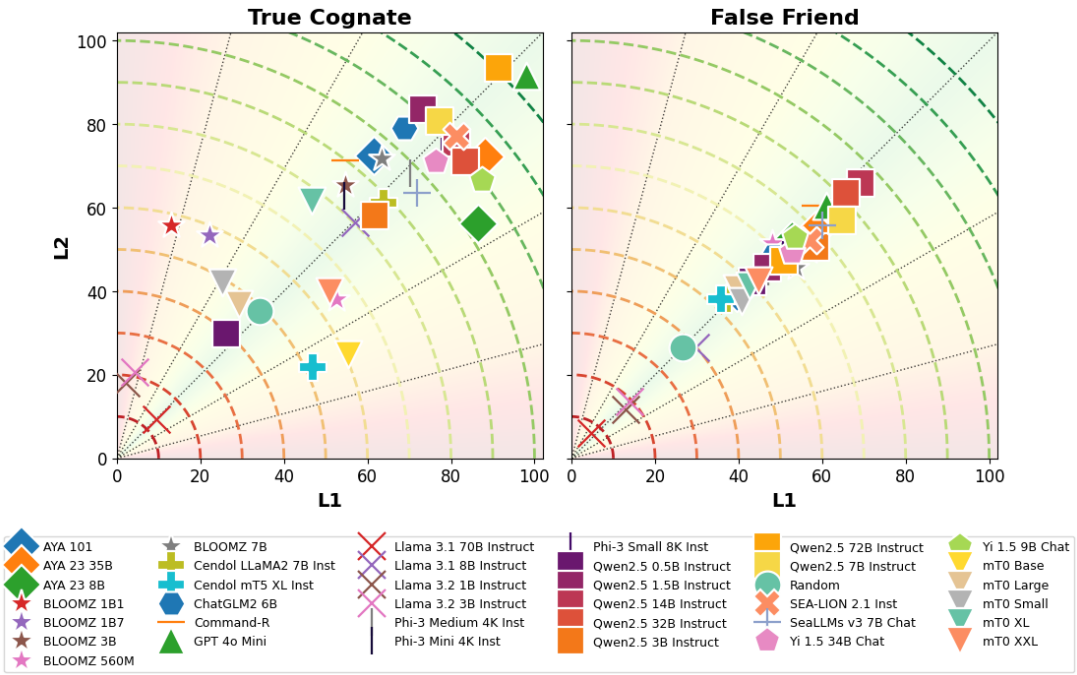


Figure 4: Stingray plot showcasing the performance of each LLM averaged across all language pairs and tasks. There is a different trend between the model performance on the **(left)** true cognate and **(right)** false friend subsets. LLMs showcase strong capability on true cognates, but close to random guessing on false friends. This highlights the inability of existing LLMs to disambiguate false friends across different languages.

4.3 Evaluation & Inference

For the inference, we conduct zero-shot prompting by prompting LLMs to answer the given prompt directly using each of the corresponding chat formats supported in each LLM. We perform two different types of inference: (1) likelihood-based inference; and (2) generation-based inference.

Likelihood-based To perform likelihood-based inference, we follow the zero-shot prompting implementation from prior works (Cahyawijaya et al., 2023b,a; Zhang et al., 2023a; Lovenia et al., 2024). For binary classification tasks, we use the label with the highest marginal likelihood given the prompt. For multiple-choice tasks, we provide the choices after the query and take the answer choice label, i.e., `A`, `B`, or `C`, with the highest likelihood. We opt for the likelihood-based for open-source LLMs as we cannot perform this on the API-based LLMs.

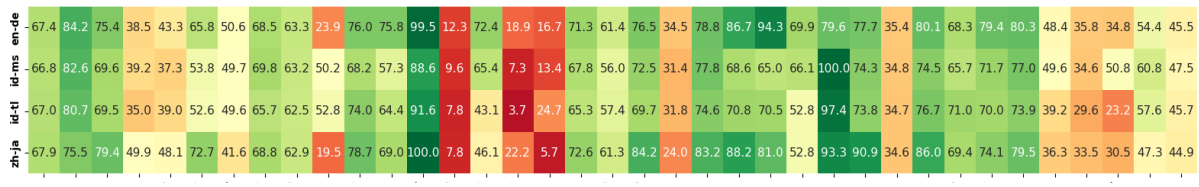
Generation-based To generalize and ensure the robustness of our results, we also do inference with a generation-based approach. The prompts are shown in 2 and with an additional sentence that asks LLMs to limit their answers to "A, B or C" or "Yes or No". To get the final result, we post-process the generated responses: as an example, for the semantic appropriateness task, LLMs sometimes answer "A and B" instead of the option "C". We test all LLMs listed in Table 2. Nonetheless, we also note that in some LLMs, they often fail to follow the given instructions.

5 Analysis and Discussion

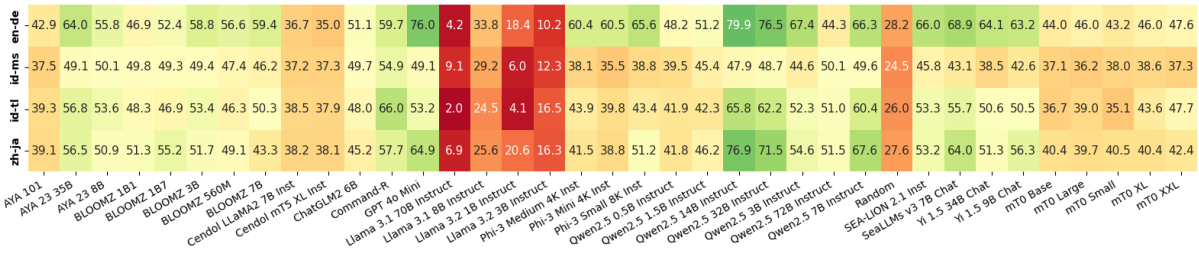
We show the stingray plot of the language-and-task aggregated results from our experiment in Figure 4. We observe a clear distinction of LLMs' cognate understanding between true cognate and false friend subsets, and provide further analysis of this behavior in the following section.

5.1 Do LLMs Understand True Cognates?

We showcase the breakdown performance per language pair on the true cognate subset in Figure 5a. Although, some smaller-scale LLMs does not perform as good, but larger LLMs tend to yield strong cognate comprehension score. Some LLMs such as Aya-23 (35B), ChatGLM2 (6B), Phi-3 Small, Qwen 2.5, Yi 1.5 (34B), and GPT-4o-mini even achieve almost perfect scores with average cognate comprehension scores $\geq$90%. This indicates that most LLMs understand the semantics of a true cognate and can incorporate it properly in both languages in the corresponding language pair.



(a) True Cognate



(b) False Friend

Figure 5: Most LLMs understand true cognates, but have limited understanding in regards to false friends in language pairs under study. We report the averaged cognate comprehension scores across the semantic correctness and usage correctness tasks.

Bias in Cognate Understanding Although achieving a high cognate comprehension score, some LLMs suffer a high cognate bias. As shown in Figure 6, LLMs such as mT0-XXL and Cendol mT5 XL show strong cognate biases towards relatively higher-resource language in the cognate language pairs including English (in English-German), Indonesian (in Indonesian-Malay and Indonesian-Tagalog), and Chinese (in Chinese-Japanese); while LLMs such as Llama-3.x, BLOOMZ, and MT0 small reflect strong cognate biases towards the other languages. This demonstrates the suitability of StingrayBench as a testbed for investigating the language selection bias in LLMs (Nomoto, 2023; Nomoto et al., 2024).

Scaling Law and True Cognate Comprehension As shown in Figure 5a and Figure 6, we clearly see the impact of the model scale in both cognate bias and cognate comprehension score. For example, BLOOMZ-560M, mT0-small, and Qwen-2.5 0.5B produce low cognate comprehension scores with a high cognate bias, while the larger scale of BLOOMZ, mT0, and Qwen-2.5 have higher cognate comprehension scores with much lower cognate bias. This signifies that the scaling law of language model (Kaplan et al., 2020; Hoffmann et al., 2022) is advantageous for understanding the sense of true cognates. However, it remains unclear why LLMs of different sizes within the same family exhibit different biases toward certain high-resource and low-resource languages and we leave this exploration for future work.

5.2 Can LLMs Distinguish False Friend?

While all LLMs show low cognate bias on the false friend subset as shown in Figure 6, most LLMs perform very poorly on the cognate comprehension score in most language pairs. For instance, most LLMs yield comprehension scores that are close to a random baseline as shown in Figure 5b. This signifies that most existing LLMs could not even distinguish the sense of false friends across different languages emphasizing an urgent need for a more advanced method on cross-lingual sense disambiguation in multilingual LLMs.

Language Representation Matters Despite the quality on the false friend subset is generally low across all LLMs, most LLMs show higher performance on the English-German language pair, including Qwen2.5 (14B and 32B), Yi (34B), Aya 23 (35B), and GPT-4o-mini. This result indicates that most of the existing multilingual LLMs, despite further tuning on other languages, are still English-centric. This observation is consistent with the fact that most existing multilingual LLMs are primarily trained on English data (Xue et al., 2020; Muennighoff et al., 2023; Chowdhery et al., 2023; Üstün et al., 2024; Aryabumi et al., 2024), highlighting the need for enhanced representation of non-English languages.

Language Similarity Affects False Friend Disambiguation We observe that the performance of the Indonesian-Tagalog language pair tends to be higher than the performance of the Indonesian-

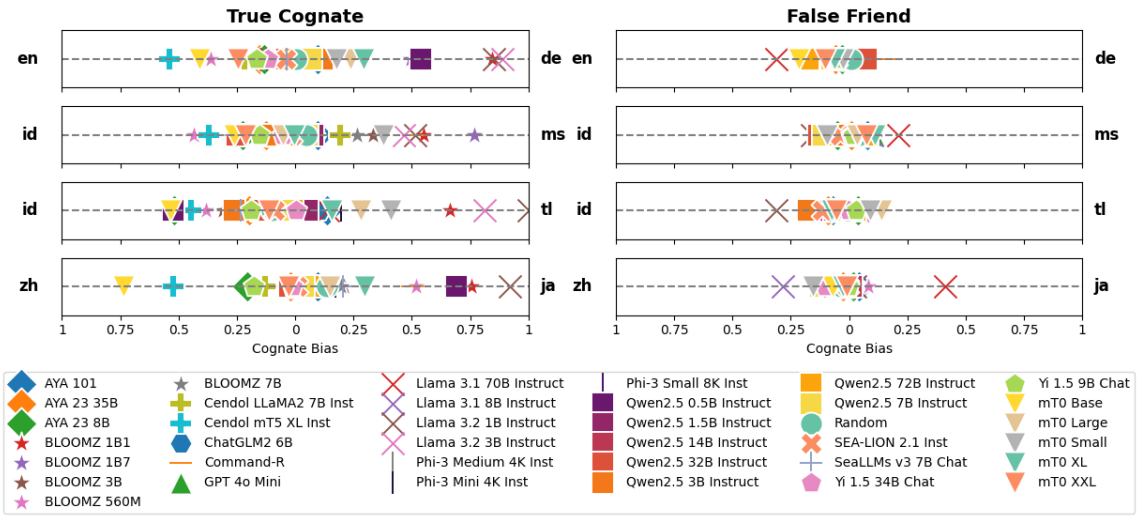


Figure 6: Cognate bias in **(left)** true cognates and **(right)** false friends for each language pair under study. We remove the sign of the cognate bias score to avoid confusion.

Malay language pair although the amount of Tagalog data is commonly lower than Malaysian data in the pretraining and supervised fine-tuning (SFT) data. For example, the mC4 corpus (Xue et al., 2020) consists of 0.21% Malay and only 0. 03% Tagalog, the pre-training corpus in PaLM (Chowdhery et al., 2023) consists of 212M Malay tokens and only 175M Tagalog, the Aya dataset (Singh et al., 2024) covers ∼4% of Malay data and <1% of Filipino (a language closely related to Tagalog) with no Tagalog data. Additionally, although we observe some positive correlation of scaling law in most subsets, we do not observe such a trend in the Indonesian-Malay subset. This signifies that existing LLMs on all scales have difficulty distinguishing false friends between these two languages.

We hypothesize that this is potentially caused by the high language similarity between Indonesian and Malay. Specifically, Indonesian and Malay fall under the same language family group (Austronesian → Malayo-Polynesian → Malayic) in both Ethnologue (Kwary and Nor Hashimah Jalaluddin, 2015; Eberhard et al., 2024) and Glotollog (Hammarström et al., 2024). Furthermore, both have great overlap in terms of lexical and grammatical aspects (Kwary and Nor Hashimah Jalaluddin, 2015; Lin et al., 2018; Nomoto et al., 2018). Some prior works (Nomoto, 2023; Nomoto et al., 2024) have also highlighted that, even a commercial LLM such as ChatGPT (Bang et al., 2023; Wu et al., 2023; Liu et al., 2023), still has the problem of differentiating between Malay and Indonesian, and often answers questions in Malay with responses in Indonesian, causing an imbalance of linguistic power, inequality between the two languages, and misrepresentation of the two languages. In this case, we can conclude that disambiguating false friends in language pairs that are highly similar, e.g., Indonesian-Malay, is a significantly more difficult problem compared to a much less similar language pair, e.g., Indonesian-Tagalog.

6 Conclusion

Our work presents a comprehensive evaluation of cross-lingual sense disambiguation in multilingual LLMs. Through the introduction of StingrayBench [5], we measure and analyze semantic understanding across languages. By studying false friends and true cognates, we have identified key factors contributing to semantic biases. Our methodology, including the stringray plot and evaluation metrics, i.e., cognate bias and cognate comprehension score, offers a novel approach to understanding cross-lingual sense disambiguation in multilingual LLMs. The generalization of our findings across various language pairs highlights the significance of this work. Our StingrayBench is not only suitable for measuring the cross-lingual sense disambiguation in LLMs, but also a suitable testbed for investigating language selection bias in multilingual LLMs. We believe that our contributions provide a foundation for further enhancing the cross-lingual capabilities of LLMs, ultimately improving their reliability and performance in diverse linguistic contexts and advancing the development of more inclusive and unbiased multilingual LLMs.

[5] Check the StingrayBench's dataset card in Appendix B

Limitation

Dataset Size Despite the enormous efforts on annotating with multiple native speakers across different language pairs, due to the limited amount of available false friends and true cognates across different language pairs, our StingrayBench consists of only around 150-200 samples per language pairs. To cater to this limitation, we try to increase the task, allowing probing of multilingual LLMs with bigger sample sizes. We leave further exploration on how to increase the amount of data of false friend and true cognate to future work.

Benchmark Coverage Due to the difficulty in finding annotators, our StingrayBench only covers four language pairs, i.e., English-German, Indonesian-Malay, Indonesian-Tagalog, and Chinese-Japanese. There are many other potential language pairs that can be covered in the benchmark, such as Sloven-Croatian, Spanish-Portuguese, etc. We expect future work to extend the generalization of our benchmark and findings to other language pairs.

Ethics Statement

This work introduces a novel benchmark for cross-lingual sense disambiguation and evaluation in multilingual large language models (LLMs), aiming to uncover biases and limitations in their semantic understanding across languages. Throughout the development of this benchmark, several ethical considerations were taken into account.

Inclusivity and Fairness The primary motivation of our work is to highlight and address the biases present in multilingual LLMs, particularly toward high-resource languages. We recognize that current language technologies often underperform speakers of low-resource languages, which could reinforce language hierarchies and contribute to the marginalization of these linguistic communities. By incorporating language pairs such as Indonesian-Malay and Indonesian-Tagalog, we strive to promote inclusivity and fairness in the evaluation of LLMs and advocate for broader linguistic diversity in NLP research.

Bias and Misrepresentation One of the key goals of our research is to identify bias in cross-lingual semantic disambiguation, especially concerning the handling of false friends and true cognates. We understand that biases in LLMs can result in misrepresentation of user intent and can have far-reaching consequences when applied in real-world scenarios. Our benchmark seeks to pinpoint these issues, providing tools for researchers and practitioners to mitigate such biases and ensure that LLMs produce more accurate and fair multilingual outputs.

Data Annotation The data used in our benchmark was carefully curated and annotated by native speakers of the respective languages to ensure linguistic accuracy and cultural sensitivity. We made every effort to fairly compensate our annotators and ensure that their contributions were recognized and valued. Additionally, we acknowledge the limitations of our dataset size and coverage and encourage further efforts to expand and diversify the benchmark in future work.

Privacy and Security Our dataset does not include any personally identifiable information or sensitive data. The false friends and true cognates were collected from publicly available resources, and no private or proprietary data were used in this research. We ensured that all data collection and usage adhered to ethical guidelines and standards in the field of natural language processing.

By addressing these ethical considerations, we aim to foster more responsible and equitable multilingual LLMs, contributing to the advancement of fair and inclusive language technologies.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet,

Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. 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, St. Julian's, Malta. Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024. Automated cognate detection as a supervised link prediction task with cognate transformer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 965–975, St. Julian's, Malta. Association for Computational Linguistics.

Keith Allan. 2009. *Concise encyclopedia of semantics*. Elsevier.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Quentin D Atkinson. 2013. The descent of words. *Proceedings of the National Academy of Sciences*, 110(11):4159–4160.

Niyati Bafna, Josef van Genabith, Cristina España-Bonet, and Z. Žabokrtský. 2022. Combining noisy semantic signals with orthographic cues: Cognate induction for the indic dialect continuum. In *Conference on Computational Natural Language Learning*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. A large and evolving cognate database. *Language Resources and Evaluation*, 56:165 – 189.

Gábor Berend. 2023. Combating the curse of multilinguality in cross-lingual wsd by aligning sparse contextualized word representations. *arXiv preprint arXiv:2307.13776*.

Michele Bevilacqua, Roberto Navigli, et al. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the conference-Association for Computational Linguistics. Meeting*, pages 2854–2864. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Samuel Cahyawijaya. 2024. Llm for everyone: Representing the underrepresented in large language models. *Preprint*, arXiv:2409.13897.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023a. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023b. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lyle Campbell. 2013. *Historical Linguistics*. Edinburgh University Press, Edinburgh.

Santiago Castro, Jairo Bonanata, and Aiala Rosá. 2018. A high coverage method for automatic false Friends detection for Spanish and Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 29–36, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Cohere For AI. 2024a. c4ai-command-r-08-2024.

Cohere For AI. 2024b. c4ai-command-r-plus-08-2024.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Liviu P. Dinu, Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023. Robocop: A comprehensive Romance BOrrowing COgnate Package and benchmark for multilingual cognate identification. In *Conference on Empirical Methods in Natural Language Processing*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World. Twenty-seventh edition.* SIL International, Dallas, Texas.

Javier Ferrando and Marta R Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. *arXiv preprint arXiv:2410.06496*.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank, editors. 2024. *Glottolog 5.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Deny Arnos Kwary and Nor Hashimah Jalaluddin. 2015. *The lexicography of Indonesian/Malay*, page 1–11. Springer, Berlin.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Nankai Lin, Sihui Fu, Shengyi Jiang, Gangqin Zhu, and Yanni Hou. 2018. Exploring lexical differences between Indonesian and Malay. In *2018 International Conference on Asian Language Processing (IALP)*, pages 178–183.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

Nikola Ljubesic and Darja Fišer. 2013. Identifying false friends between closely related languages. In *BSNLP@ACL*.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Chia Tai, Ayu Purwarianti, Sebastian Ruder, William Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng-Xin Yong, and Samuel Cahyawijaya.

2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. *Preprint*, arXiv:2406.10118.

Mohd Sharifudin bin Yusop and Mahyudin Al Mudra. 2015. *Kamus Komunikatif Nusantara: Indonesia-Malaysia, Malaysia-Indonesia*. Balai Kajian dan Pengembangan Budaya Melayu, Yogyakarta.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Hiroki Nomoto. 2023. Issues surrounding the use of ChatGPT in similar languages: The case of Malay and Indonesian. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 76–82. Association for Computational Linguistics.

Hiroki Nomoto, Shiro Akasegawa, and Asako Shiohara. 2018. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA*, 65:47–66.

Hiroki Nomoto, David Moeljadi, and Farhan Athirah Abdul Razak. 2024. Masalah teknologi dan isu sosial berkaitan penggunaan ChatGPT dalam bahasa Melayu [Technological and social issues related to using ChatGPT in Malay]. *RENTAS: Jurnal Bahasa, Sastera dan Budaya*, 3(1):1–22.

David Ong and Peerat Limkonchotiwat. 2023. SEA-LION (Southeast Asian languages in one network): A family of Southeast Asian language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 245–245, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex

Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Tommaso Pasini. 2021. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4936–4942.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206.

Rusdi Abdullah. 2016. *Kamus Kata Bahasa Melayu Malaysia-Bahasa Indonesia*. Penerbit Universiti Kebangsaan Malaysia, Bangi.

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, Bangkok, Thailand. Association for Computational Linguistics.

Ying Su, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. Multilingual word sense disambiguation with unified sense representation. *arXiv preprint arXiv:2210.07447*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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. *Preprint*, arXiv:2307.09288.

Ana Sabina Uban and Liviu P Dinu. 2020. Automatically building a multilingual lexicon of false friends with no supervision. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3001–3007.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction fine-tuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Warren Weaver. 1949. Translation. In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA. Reprinted from a memorandum written by Weaver in 1949.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023a. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Ruochen Zhang and Carsten Eickhoff. 2024. CroCoSum: A benchmark dataset for cross-lingual code-switched summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4113–4126, Torino, Italia. ELRA and ICCL.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2024. The same but different: Structural similarities and differences in multilingual language modeling.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023b. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A Annotation Guideline

A.1 Annotation Objective

The goal of this annotation task is to create a dataset that distinguishes between false friends and true cognates across various language pairs. Annotators will work with native speakers to identify and categorize common words, construct sentences, and translate them to ensure accurate representation.

A.2 Word Selection and Criteria

False Friends: Words with the same spelling or characters but different meanings in the respective languages. **True Cognates**: Words with the same spelling, characters, and meanings in both languages. For collecting the common words, annotators incorporate the following sources:

- Wiktionary's lists of false friends [6]
- Kamus Komunikatif Nusantara: Indonesia-Malaysia, Malaysia-Indonesia (Mohd Sharifudin bin Yusop and Al Mudra, 2015)
- Kamus Kata: Bahasa Melayu Malaysia-Bahasa Indonesia (Rusdi Abdullah, 2016)

A.3 Annotation Process

Annotation Flow

- **Translation and Replacement** For false friends, given a false friend word, an annotator will make the correct sentence in their language and translate it into English. The annotator of the other language in the pair will then translate the English translation into their native language. The target false friend word will be replaced with a different word that conveys the intended meaning. This will result in a semantically odd sentence.
- **Cognate Agreement and Translation**: For true cognates, both annotators will first agree on an English sentence. The English sentence will then be translated into their respective native languages to construct the true cognate sentence pair.

Allowed Variations

- English-German: Words differ in capitalization with a maximum of one edit distance.
- Chinese-Japanese: Words with different characters developed from the same origin.
- Indonesian-Malay: Exact match words.
- Indonesian-Tagalog: Words with maximum of one edit distance.

A.4 Dataset Entries

- **False Friends**: Each false friend will typically have two entries — one for the correct usage and one for the incorrect usage.
- **True Cognates**: Each true cognate will have one entry, as both native language sentences translate correctly.

A.5 Examples

False Friend (Indonesian-Tagalog):

- Indonesian sentence: Pagi ini saya begitu senang
- English translation: This morning, I feel so happy
- Tagalog translation: Ako ay masaya ngayong umaga
- Replacement: Ako ay masaya ngayong pagi ("Today's stingray, I feel happy")

True Cognate (Indonesian-Malay)

- English sentence: "That apple has many caterpillars."
- Indonesian sentence: Apel itu banyak ulat
- Malay sentence: Epal itu ada banyak ulat.

A.6 Additional Guidance for Annotators

- Ensure a clear understanding of the word's meaning and context.
- Construct sentences that are natural and grammatically correct in your native language.
- Pay attention to the nuances and potential variations in word usage.
- For false friends, aim for a semantically odd translation to highlight the semantic differences between the two sentences.
- Collaborate effectively with your partner annotator to ensure accurate translations and representations.

B Dataset Card

Dataset Name: StingrayBench

B.1 Dataset Description

Overview StingrayBench is a dataset designed to evaluate models' understanding of semantic appropriateness and cognate word usage across multiple language pairs. The dataset focuses on false friends and true cognates, which are words with similar

[6] https://en.wiktionary.org/wiki/Category:False_cognates_and_false_friends

spellings or characters but different meanings or additional meanings in different languages.

Language Pairs The dataset covers the following language pairs:

- English-German (EN-DE)
- Chinese-Japanese (ZH-JA)
- Indonesian-Malay (ID-MS)
- Indonesian-Tagalog (ID-TL)

B.2 Dataset Construction

Native speakers of the respective languages were involved in constructing the dataset. They listed common words between language pairs and their meanings, consulting resources on false friends. The words were then categorized as false friends or true cognates.For each word, annotators created sentences in their native language and provided English translations. The sentences were designed to showcase the correct and incorrect usage of the target words. In the case of false friends, the sentences were manipulated to produce semantically odd translations.

B.3 Dataset Statistics

The dataset contains a total of 705 entries, including: 259 true cognate entries and 446 false friend entries. The distribution of entries across language pairs is as follows:

- EN-DE: 196 entries (98 true cognates, 98 false friends)
- ZH-JA: 165 entries (51 true cognates, 114 false friends)
- ID-MS: 186 entries (52 true cognates, 134 false friends)
- ID-TL: 158 entries (58 true cognates, 100 false friends)

B.4 Task Formulation

Semantic Appropriateness In this task, models are prompted to determine which sentence is more semantically appropriate. The prompt includes two sentences from the language pair and a third option indicating that both sentences are appropriate. This task aims to test the model's comprehension and understanding of the semantic nuances between the language pairs.

Usage Correction The usage correction task focuses on the correct usage of specific cognate words. Models are prompted with a sentence containing a cognate word and asked to determine if the word's usage is correct. This task provides a more targeted evaluation of the model's ability to handle cognate words accurately.

B.5 Example Prompts and Completions

Semantic Appropriateness .

Prompt:
Which sentence is more semantically appropriate?
A. "Ich habe einen Arm." (German)
B. "I have an Arm." (English)
C. "Both sentences are appropriate."

Target Completion: "C. Both sentences are appropriate."

Usage Correction .

Prompt:
Is the usage of "pagi" in this sentence correct?
"Ako ay masaya ngayong pagi." (Tagalog)

Target Completion:
"No, the usage of 'pagi' is incorrect. 'Pagi' means 'stingray' in Tagalog, and the sentence should use 'umaga' for 'morning'."

B.6 Dataset Licensing Information

To promote accessibility, encourage collaboration, and facilitate knowledge sharing, StingrayBench will be made available to the public under the Creative Commons Attribution-ShareAlike 4.0 International (CC-BY-SA 4.0) license. This license ensures that the dataset is accessible and can be utilized by a wide range of individuals and organizations including for commercial users.