Predicting Multilingual Classification and Translation Performance of LLMs with Cross-Lingual Alignment – Is English Enough?

Adnan Al Ali[1], Kathy Hämmerl[2,3], Jindřich Libovický[1], Alexander Fraser[2,3]
[1]Charles University, Faculty of Mathematics and Physics, Czech Republic
[2]Technical University of Munich, Germany
[3]Munich Center for Machine Learning
Correspondence: alali@ufal.mff.cuni.cz

Abstract

Multilingual large language models (LLMs) have been shown to perform better on non-English classification tasks when the representations of the given language are more aligned to English within the model. Several cross-lingual alignment (CLA) scores have been proposed for use with LLMs, along with multiple approaches for extracting embeddings from the models. We provide a comparative analysis of 27 CLA score variants, examining how they differ and how well each predicts downstream performance across three tasks. Crucially, while LLMs are widely used for generative tasks such as machine translation, prior work has focused almost exclusively on classification. We therefore investigate whether CLA scores are similarly predictive of translation performance. To enable computing correlations across target languages, we propose a PMI-based translation metric, which is less dependent on the target language and correlates strongly with CHRF. We find that CLA with English predicts translation quality comparably to or better than source–target CLA, providing new evidence that LLMs use English as an internal pivot language.

1 Introduction

Multilingual large language models (LLMs) exhibit remarkable multilingual capabilities (Singh et al., 2025) in tasks such as machine translation (Goyal et al., 2022), reading comprehension (Bandarkar et al., 2024), and text classification (Adelani et al., 2024), among others. While these capabilities have been extensively benchmarked, we have little understanding of how multilinguality works.

Cross-lingual alignment (CLA) scores can reliably predict LLM performance on non-English downstream tasks without needing to explicitly run the tasks, as Ravisankar et al. (2026) and Kargaran et al. (2025) show. They specifically measure alignment scores with English, based on the claim that LLMs implicitly translate prompts to English in their hidden states (Wendler et al., 2024). Alignment scores can be obtained from embeddings (Kargaran et al., 2025; Del and Fishel, 2022; Artetxe and Schwenk, 2019, *inter alia)* or tokenisation (Hämmerl et al., 2025; Limisiewicz et al., 2023).

In this work, we comprehensively review existing methods for obtaining sentence-embedding-based CLA scores, including sentence embedding extraction. We test the correlation between the scores and performance on three tasks: (1) *SIB-200* (text categorisation; Adelani et al., 2024), (2) *Belebele* (reading comprehension; Bandarkar et al., 2024), and (3) *Flores-200* & *BOUQuET* (machine translation; NLLB Team et al., 2022; Andrews et al., 2025).

CLA has been studied extensively on encoder-only models, but only a limited number of recent papers focus on decoder-only models (Hämmerl et al., 2025; Ravisankar et al., 2026; Kargaran et al., 2025), and none address the translation task, which remains understudied despite its popularity.

Therefore, our work concentrates on translation. To our knowledge, no prior work has inspected the impact of CLA scores on translation performance. We aim to answer an essential question: when translating from language `src` to `tgt`, which CLA is more predictive of quality: (1) `src-tgt`, (2) `src-en`, or (3) `en-tgt`? To answer this, we propose a translation evaluation metric based on pointwise mutual information (PMI) which performs similarly to CHRF but depends less on the target language – a property crucial to our analysis.

We analyse the contributions of CLA scores to prediction by finding optimal weights in an ensemble of the different alignment directions. Our findings suggest that CLA scores with English (`src-en` and `en-tgt`) are generally more predictive, contributing new evidence for the claim that LLMs use English as a pivot language from a yet unexplored perspective.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 3719–3741
October 24-29, 2026 ©2026 Association for Computational Linguistics

The main contributions of this work are: 1: A comprehensive review of existing methods for measuring CLA scores from sentence embeddings. 2: A new multilingual **fewshot** method for extracting sentence embeddings from LLMs (Section 3.1). 3: Expanding the study of CLA in LLMs to machine translation (Section 6). 4: A new translation evaluation metric that is less target-language dependent (Section 6.1). 5: New evidence for the claim that LLMs use English as an internal pivot language. We publish our code on GitHub.[1]

2 Background and Related Work

Cross-lingual alignment. CLA, in the broad sense, is 'the similar representation of similar meanings regardless of the language' (Hämmerl et al., 2024). Previous work has used various methods of measuring this concept to estimate zero-shot cross-lingual transfer in encoder-only models (Conneau et al., 2020; Del and Fishel, 2022; Limisiewicz et al., 2023) and decoder-only performance based on alignment with English (Kargaran et al., 2025; Ravisankar et al., 2026). While we focus on the latter, methods developed for encoder models can be adapted to decoders with minimal modifications.

Embedding-based CLA has also been studied in encoder-decoder machine translation. Kudugunta et al. (2019) and Qu et al. (2025) used the similarities of the embeddings to show that when the target language is known, the encoder will map the input sentence into that language's space (as opposed to a language-independent space). Compared to that, the sentence embeddings obtained from an LLM are completely agnostic to the target language or the translation task as a whole, which considerably changes the predictive properties of the alignment.

Obtaining text embeddings. Embedding-level alignment scores (described below) require a fixed-length vector representation for each sentence.[2] In contrast, LLMs assign a hidden state (embedding) to individual tokens. Various approaches for pooling token embeddings into a single vector exist; we refer to these as *sentence representations*.

Taking the **mean** of the embeddings is a straightforward option; however, Jiang et al. (2022) found that this approach with BERT (Devlin et al., 2019) performs worse than averaging static embeddings for semantic similarity tasks.

In decoder-only Transformers (Vaswani et al., 2017), attention is masked so that each token can attend only to previous tokens, meaning later tokens carry more information. Position-weighted mean (**p-wm**; Muennighoff, 2022) leverages this by computing a weighted mean of the embeddings with weights proportional to position index. Alternatively, one can simply use the **last** token embedding.

To better aggregate sentence meaning, Jiang et al. (2022, 2024) propose the **prompt** `This sentence: "[text]" means in one word: "`. However, this approach would require a human translation of the prompt into every language considered, which is difficult to obtain; otherwise, the embeddings would implicitly reflect alignment between English and the inspected language.

For completeness, we use all sentence representations listed here in our experiments, including the English **prompt**, and extend this list with our novel approach in Section 3.1.

Embedding-level alignment scores. Consider a parallel dataset of D sentences in the source and target languages, represented as d-dimensional vectors. Suppose the embeddings are stored in matrices $S \in \mathbb{R}^{d \times D}$ (source) and $T \in \mathbb{R}^{d \times D}$ (target). We consider the following metrics.

The most straightforward approach, average **cos**ine similarity, is the per-sentence-pair average of

$$\cos(S_{*i}, T_{*i}) = \frac{S_{*i} T_{*i}^{\top}}{\|S_{*i}\| \|T_{*i}\|}, \tag{1}$$

However, Ethayarajh (2019) noted that embeddings occupy an anisotropic space, making direct comparisons unreliable.

xSIM (Artetxe and Schwenk, 2019) uses cosine similarity indirectly by computing accuracy in parallel corpus retrieval. In its simplest form **abs**, the retrieval similarity between two embeddings is their cosine similarity $\cos(S_{*i}, T_{*i})$. To mitigate scale inconsistencies, the **dist** and **ratio** variants use a margin between the cosine similarity of S_{*i} and T_{*i} and the average cosine similarity of their k nearest neighbours within their respective languages:

$$\begin{aligned} \mathrm{margin}\Big(&\cos(S_{*i}, T_{*i}), \\ &\frac{1}{2k}\Big(\sum\nolimits_{S_{*j} \in \mathrm{kNN}(S_{*i})} \cos(S_{*i}, S_{*j}) \\ &\quad + \sum\nolimits_{T_{*j} \in \mathrm{kNN}(T_{*i})} \cos(T_{*i}, T_{*j})\Big)\Big) \end{aligned} \tag{2}$$

[1] https://github.com/KathyHaem/cla-metrics

[2] While we work with sentences in the traditional linguistic sense, any reasonably long chunk of text can be considered a sentence in this context.

where margin $\mathrm{margin}(a, b)$ is either $a - b$ (for **dist**) or a/b (for **ratio**).

Dali$_{\text{weak}}$ (Ravisankar et al., 2026) computes retrieval accuracy on a corpus where each sample consists of a context c and completions $a_1, \ldots, a_n$, of which exactly one is correct (a^*). Retrieval is successful if the correct source-language context-completion pair has greater cosine similarity to its target-language translation than to any incorrect target-language completion:

$$\arg\max_{a^{\text{tgt}}} \big(\cos(c^{\text{src}} a^{*\text{src}}, c^{\text{tgt}} a^{\text{tgt}})\big) = a^{*\text{tgt}}. \quad (3)$$

In our experiments, we use Belebele (Bandarkar et al., 2024) as the dataset of contexts and completions.

Finally, average neuron-wise correlation (**ANC**; Del and Fishel, 2022) leverages the other dimension of the S and T matrices; for each position ('neuron') $j \in \{1, \ldots, d\}$ in the embedding vector, calculates the Pearson correlation coefficient over the j-th row of the dataset $\rho(S_{j*}, T_{j*})$ and returns the per-neuron average, rather than the working with the individual sentence embeddings stored in the columns.

We test all the listed methods in our experiments.

Tokeniser-level alignment scores. Vocabulary overlap between languages has long been considered a factor in cross-lingual transfer (e.g., Wu and Dredze, 2019). Rather than simply counting shared tokens, Limisiewicz et al. (2023) use the similarity of two languages' token distributions to predict downstream transfer performance on certain tasks. However, very few tokens are shared across script boundaries.

Hämmerl et al. (2025) show that subword token *alignability* is more predictive of downstream transfer than literal subword token distribution overlap. They define token alignability via the log-likelihood score of the **Eflomal** word aligner (Östling and Tiedemann, 2016), averaged over all sentence pairs in a parallel corpus. The eflomal score acts as a proxy for how well the tokeniser supports learning shared semantics across languages, and operates quite differently from embedding-level alignment scores.

English as a pivot language. Contemporary LLMs are trained predominantly on English data (OpenAI et al., 2024), yet perform well in other languages. Wendler et al. (2024) investigated this by inspecting the hidden layers of Llama 2 (Touvron et al., 2023), whose pre-training data is 89.7% English. They prompted the model to translate a word between two non-English languages, decoded the intermediate hidden states, and found that the resulting tokens were often the English equivalent of the translated word, suggesting that the model operates internally in English even when prompted in another language. We provide additional evidence for this claim from a different perspective using novel methodology.

Kargaran et al. (2025) and Ravisankar et al. (2026) build on this claim, showing that retrieval-based CLA between a given language and English correlates well with monolingual performance in that language on tasks such as reading comprehension, language understanding, and narrative understanding.

Explicitly using a high-resource language, such as English, as a pivot has been a common practice in machine translation. The techniques include translating into the pivot language as an intermediate step, generating pivot-based synthetic data, or transferring a model trained on the pivot language (Kim et al., 2019). Evidence presented above suggests that this approach emerges in LLMs as well. We aim to further study this behaviour using CLA.

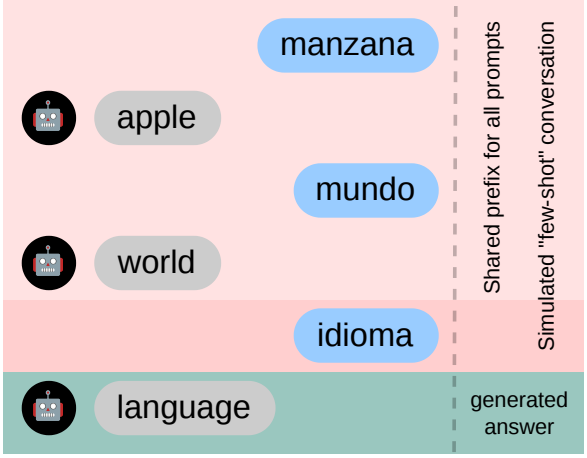

Figure 1: Example two-shot prompting conversation for translation from Spanish to English.

Few-shot prompting. Few-shot prompting involves presenting an LLM with multiple solved examples of a task, followed by an unsolved instance. The model then follows the previous pattern and generates the answer for that instance. This approach is typical for non-instruction-tuned models (Brown et al., 2020). In some of our experiments,

we opt for a few-shot approach and exploit its absence of an explicit instruction, since translating a prompt into all languages would be unfeasible, and using a shared English prompt would implicitly embed the alignment with English.

We further expand the few-shot learning approach to instruction-tuned LLMs, where we apply a chat template to the traditional few-shot prompt to simulate a conversation without an explicit instruction. Figure 1 shows an example few-shot prompt for translation.

3 Experimental Setup

In our experiments, we use five methods to represent sentences as fixed-length vectors and six methods to measure CLA scores from those representations, together with a tokeniser-level CLA metric, yielding 27 feasible combinations in total. We examine how these combinations correlate with performance on downstream tasks.

We conduct our experiments with the following models in their base and instruction-tuned versions: `Qwen3-14B[-Base]` (Yang et al., 2025), `gemma-3-12b-{pt,it}` (Gemma Team et al., 2025), and `Ministral-3--14B-{Base,Instruct}-2512` (Liu et al., 2026). To ensure a fair comparison, we disable the chain-of-thought phase for all models. In all of the experiments, we work with 44 languages of various families, resource-richnesses, and writing systems, listed in Appendix A. Throughout, we use the Pearson correlation coefficient as the measure of correlation.

3.1 Computing CLA Scores

We compute all alignments on the 44-language subset of the *Flores-200*-`dev` dataset (Goyal et al., 2022; 997 samples). For the embedding-level CLA scores, we use the sentence representation approaches described in Section 2: **mean**, **p-wm** (position-weighted mean; Muennighoff, 2022), **last** token, and English **prompt** (Jiang et al., 2024).

Additionally, *we propose a new embedding extraction method:* **fewshot**, which guides the LLM to concentrate the meaning of a sentence into a single token. This is similar to the **prompt** method (Jiang et al., 2024), but **fewshot** relies neither on an English prompt – which carries implicit information about alignment with English – nor on translating the prompt into each language.

The **fewshot** approach is inspired by *DefSent* (Tsukagoshi et al., 2021), who fine-tuned a BERT model to predict a masked word following its definition. Then, they used the embeddings of the `[MASK]` token preceded by a sentence to embed. Unlike in DefSent, we do not fine-tune the model; we prompt it with a few-shot prompt to predict an entity from its description. We use the sentence to be embedded as the last description and extract embeddings from the token preceding the predicted entity. As the source of our few-shot data, for each language, we take the first five Wikidata entities (Vrandečić and Krötzsch, 2014), sorted by QID, with descriptions of at least 10 characters. Listing 1 in Appendix B shows an example English prompt.

Since all alignment scores are calculated per layer, we select the layer with the *maximum* alignment score, following Kargaran et al. (2025). Next, we pair the sentence representation approaches with the CLA metrics described in Section 2: **cos**ine similarity, *xSIM* scores (**abs**, **dist**, **ratio**; Artetxe and Schwenk, 2019), **Dali**[3] (Ravisankar et al., 2026), and **ANC** (Del and Fishel, 2022).

For tokeniser-level CLA scores, we follow Hämmerl et al. (2025) and use the **Eflo**mal score (Östling and Tiedemann, 2016) to estimate token alignability in the parallel corpora. Unlike Hämmerl et al. (2025), we do *not* estimate the priors on a larger dataset – due to data scarcity – and instead compute the score directly from *Flores-200*-`dev`. Since a lower **Eflo**mal score indicates better alignability, we negate it to obtain positive expected correlations with task performance.

Correlations within the alignment scores. The alignment of a given language to English is strongly correlated with its alignment to other non-English languages. To demonstrate this, we correlate each `src` language's alignment to English with its average alignment to all non-English languages (`src-[~en]`) in Table 1 (for `Qwen3`). This finding implies that alignments cannot be viewed as complementary in an ensemble combining multiple alignment directions (such as Equation 7): even if both `src-en` and `src-tgt` CLAs correlate with translation quality, their combination may not yield a higher correlation. This means that it is not sufficient to observe the correlations individually; instead, we focus on how much one alignment direction can improve others in the translation task

[3] We only pair **Dali** with the **last** token sentence representation, following Ravisankar et al. (2026). Moreover, **Dali** scores are estimated on the Belebele dataset, instead of Flores.

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	95.3	63.3	87.7	87.7	87.7	–	
last	97.6	81.5	89.6	89.6	89.7	98.4	
p-wm	59.8	91.5	95.2	95.2	95.2	–	58.0
mean	97.7	89.6	55.3	55.3	55.4	–	
prompt	–	91.6	99.6	99.6	99.6	–	

Table 1: Strong correlation measured across the `src` languages between the `src-en` and `src-[~en]` alignment scores for Qwen3-14B. Values in per cent.

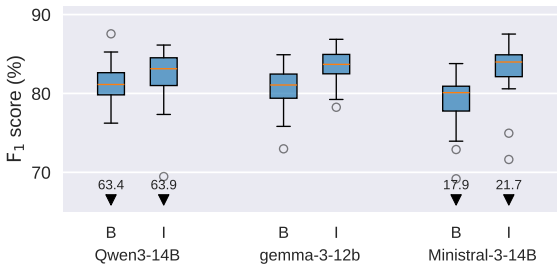

Figure 2: Macro F_1 score distribution across languages for **SIB-200**. Cropped outliers are marked by ▼.

(Section 6).

4 SIB-200 Benchmark

The task in *SIB-200* (Adelani et al., 2024) is to categorise a text segment into one of seven topics. As the category is in English, we use an English prompt, stated in Listings 2 (base models) and 6 (instruction-tuned models) in Appendix B. We compute the macro-averaged F_1 score on the `train` split (701 samples) and correlate it with the `src-en` CLA score across each `src` language, reporting the per-language correlation for each CLA metric and representation.

4.1 Results and Discussion

Overall, the models achieved satisfactory performance on SIB-200, as shown in Figure 2. `Ministral-3` achieved the lowest score on Amharic, indicating its set of supported languages is more limited than the other models.

As Table 2 shows, `src-en` CLA metrics correlate moderately to strongly with performance, with **prompt** repeatedly appearing as the best-performing sentence representation. This is likely because, in this task, we prompt the models in English, and the **prompt** representation implicitly encodes understanding of English prompts. We show detailed results in Table 8 in Appendix D.

	Qwen3		gemma-3		Ministral-3	
	base	inst.	base	inst.	base	inst.
corr.	81.8	85.9	70.9	67.6	**97.5**	97.4
repr.	prompt	prompt	p-wm	mean	prompt	prompt
metr.	abs	abs	cos	abs	abs	abs

Table 2: Best correlation measured across the `src` languages of `src-en` alignment score and the **SIB-200** F_1 score in the `src` language for all models. Values in per cent.

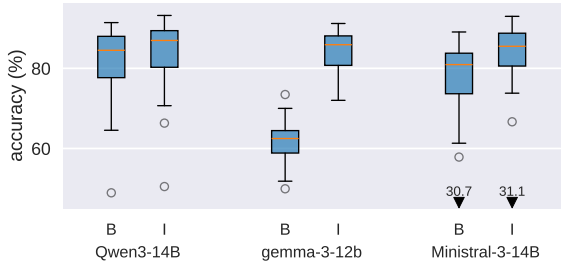

Figure 3: Accuracy distribution across languages for **Belebele**. Cropped outliers are marked by ▼.

5 Belebele Benchmark

Belebele (Bandarkar et al., 2024; 900 samples) is a multiple-choice reading comprehension dataset where the task is to select the correct answer from four options given a context and a question. Each part (context, question, options) is translated into each supported language, yielding a fully monolingual task in each language (unlike SIB-200). To avoid using English in the prompt, we use a few-shot prompting approach that shows three answered examples before the sample to be solved.

The prompting template adds only newline characters to separate the parts of the text (or role switching in the case of instruction-tuned model). We do *not* introduce option numbering or field names (such as 'Context:'). The LLM answers by explicitly stating one of the options, without sampling during generation. To save resources, we limit generated tokens to the smallest prefix that uniquely distinguishes each answer and check whether the generated prefix matches that of the correct answer. Listings 3 and 4 in Appendix B show an example prompt for base and instruction-tuned models, respectively.

We measure prediction accuracy for each language and inspect the correlation between accuracy and the `src-en` CLA score over each `src` language.

	Qwen3		gemma-3		Ministral-3	
	base	inst.	base	inst.	base	inst.
corr.	91.7	95.2	70.0	85.0	**97.4**	95.6
repr.	prompt	prompt	p-wm	p-wm	p-wm	fewshot
metr.	ANC	ANC	abs	ANC	ANC	abs

Table 3: Best correlation measured across the `src` languages of `src-en` alignment score and the **Belebele** accuracy in the `src` language for all models. Values in per cent.

5.1 Results and Discussion

Figure 3 shows that the models again achieved solid results, with multiple models exceeding 80% accuracy on average. As with SIB-200, `Ministral-3` achieved the lowest score on Amharic.

As Table 3 shows, the correlation between `src-en` CLA metrics and Belebele performance is moderate to strong, with **p-wm** repeatedly appearing as the optimal representation and **ANC** as the optimal CLA metric. Detailed correlations per model are in Table 9 in Appendix D.

6 Translation Benchmark

Unlike the previous monolingual tasks, translation involves two languages: `src` and `tgt`. A natural question is whether the `src-tgt` CLA score predicts translation quality well. Alternatively, if LLMs use English as a pivot language, `src-en` and `en-tgt` CLA may correlate more strongly with translation quality.

We primarily use the `devtest` split of *Flores-200* for translation (1011 samples) and additionally the `dev` split of the *BOUQuET* dataset (Andrews et al., 2025) for out-of-domain validation. We consider all the possible pairs of the 44 languages in both directions. To avoid English prompts that might bias alignment measurements, we use a few-shot prompt of five examples, followed by the sentence to translate. An example is in Listing 5, Appendix B. We generate translations without sampling, limiting the number of tokens to 1.5 times the length of the reference.

We evaluate translations using the CHRF score[4] (Popović, 2015) at the system level for each language pair. A limitation of CHRF is that its value depends strongly on the target language's writing system, as characters carry different amounts of information across writing systems.

This restricts us to measuring correlations within a fixed target language and averaging across target languages. Thus, we can assess the performance of `src-tgt` and `src-en` CLA scores, but assessing `en-tgt` is impossible in this setting, as it is constant under a fixed target language. We tackle this problem in the following section.

6.1 PMI as a Translation Quality Metric

To overcome the target-language sensitivity of CHRF, we propose a task-specific machine translation metric: point-wise mutual information (PMI) between the source sentence and its translation.

We aim to reduce target-language dependence by normalising the likelihood of a translation given the source sentence ($p(y|x)$) by its unconditional (prior) likelihood ($p(y)$) and compute the logarithm of the value:

$$\log \frac{p(y|x)}{p(y)} = \log \frac{p(x,y)}{p(x)p(y)} = \mathrm{PMI}(x,y) \quad (4)$$

We compute the conditional log-likelihood using the same prompt as for generating translations. For the prior log-likelihood, we use a prompt with the same five sentences but only in the target language to introduce the domain and language. An example is in Listing 7, Appendix B. In both cases, we concatenate the few-shot prefix (m tokens) and the translation (n tokens) into a single string s and compute the length-normalised log-likelihood of the translation:

$$\log p(\cdot) = \frac{1}{n} \sum_{i=m+1}^{m+n} \log p(s_i|s_1, \ldots, s_{i-1}) \quad (5)$$

We validate this metric via its strong Pearson correlation with CHRF, measured within a fixed target language, shown in Table 4. The correlation is particularly strong for base models. A notable outlier is `gemma-3` instr. on the Flores dataset. A possible explanation is training data contamination, which could skew the prior distribution.

Intuitively, PMI should achieve target-language independence, as all the language-specific factors that influence the conditional probability of the translation are also included in the prior probability, and the metric measures the relative increase in probability between the two. We empirically discuss target-language independence in Appendix C.

Note that we propose PMI for our specific experimental setup, rather than as a translation evaluation metric for general use, as its value is unbounded and may vary across models. Nonetheless, it may

[4]Using default parameters: `ncorder=6`, `nworder=2`, β=2.

	Qwen3		gemma-3		Ministral-3	
	base	inst.	base	inst.	base	inst.
flores	91.0	77.6	**96.3**	26.2	95.8	65.8
bouquet	95.0	83.0	96.0	78.5	95.8	77.7

Table 4: Correlations between system-level CHRF and PMI, measured within a fixed target language and then averaged across the target languages. Values in per cent.

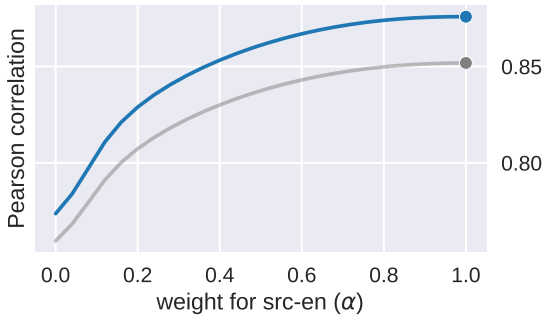

Figure 4: Per-target-language average Pearson correlation between translation CHRF and the α-weighted convex combination of `src-en` `src-tgt` CLA scores for **Ministral-3 base**, **ANC**, **mean** representations. **Blue** plot: Flores, **grey** plot: BOUQuET.

be used in any experiments in which the exact value is not important, only the correlation within a fixed model.

6.2 Analysis with CHRF

Table 5 shows CHRF scores for selected target languages; the complete lists are in Tables 10 (Flores) and 11 (BOUQuET), Appendix D. For all target languages, `gemma-3` achieved the highest results. The scores for `Ministral-3`-instruct are low because it frequently failed to follow the few-shot template and provide a translation. As noted above, performance across target languages is not directly comparable: e.g., CHRF is lower for Chinese than for other high-resource languages, because Chinese characters carry more information per character.

	Qwen3		gemma-3		Ministral-3	
	base	inst.	base	inst.	base	inst.
en	59.4	58.4	**62.4**	60.0	59.4	45.6
it	46.0	46.0	49.7	**50.1**	47.1	18.7
ar	39.1	38.0	44.2	**44.7**	40.3	21.7
zh	19.5	19.3	**21.5**	**21.4**	20.5	8.7
am	7.9	8.8	**24.5**	20.6	4.6	6.5

Table 5: Average **Flores** translation performance (CHRF; over all source languages) for the displayed target languages. Values in cent.

Correlations with `src-en` and `src-tgt` are shown in Table 6 top and middle (overview) and Table 12, Appendix D (per model). The **ANC** metric repeatedly achieves the best results across models. There is no dominant sentence representation that wins consistently across all models. In most cases, `src-en` CLA correlates more strongly with CHRF than `src-tgt` does. Given the correlations between CLA metrics described in Section 3.1, we next ask whether `src-tgt` carries any association beyond that contained in `src-en`.

To assess this, we measure the improvement in correlation from a convex combination of the two. For each sentence representation and CLA metric pair, we find the optimal convex combination weight α, maximising the correlation of

$$\alpha \mathrm{CLA}_{\texttt{src-en}} + (1-\alpha)\mathrm{CLA}_{\texttt{src-tgt}} \quad (6)$$

with CHRF averaged across target languages. We search over $[0, 1]$ with step size $1/25$.

The lower part of Table 6 shows the correlations with this optimal convex combination. Detailed results are in Tables 13 and 16, Appendix D. Although both `src-en` and `src-tgt` correlate well with CHRF, combining them increases the correlation only slightly. Figure 4 illustrates the relationship between α and the correlation for Ministral-3 base, the model with the strongest overall correlation. The BOUQuET validation shows that the correlation is retained even when the alignment and translation datasets come from different domains.

6.3 Analysis with PMI

Unlike with CHRF, we do not fix the target language when analysing PMI. Instead, we compare all language pairs simultaneously. In this setting, using only one alignment from `src-en` or `en-tgt` is impractical, as it would yield duplicate alignment scores.[5] Therefore, we omit the single-alignment experiment and proceed directly to analysing the convex combination correlation. This time, we have three weights, with two degrees of freedom: $\alpha \in [0, 1], \beta \in [0, 1-\alpha]$:

$$\alpha \mathrm{CLA}_{\texttt{src-en}} + \beta \mathrm{CLA}_{\texttt{en-tgt}} + (1-\alpha-\beta)\mathrm{CLA}_{\texttt{src-tgt}} \quad (7)$$

We show the overview of the results of this experiment in Table 7 (top) and the details in Tables 14, 15 and 17, Appendix D. Figure 6 illustrates the grid

[5] For example, if we used only `src-en` CLA scores, the pairs `it-ar` and `it-pt` would both be assigned the same `it-en` CLA score.

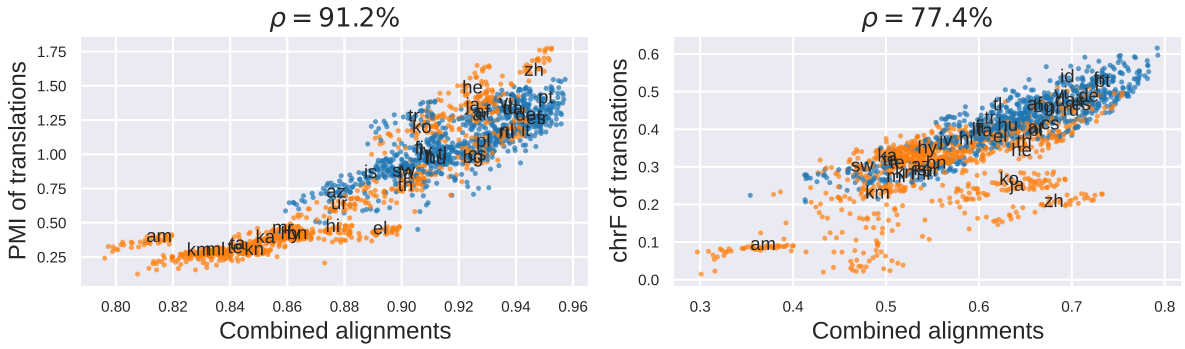

Figure 5: CHRF (right) clusters target languages by character information density, while clusters are considerably less pronounced for PMI (left). Each point is a language pair; non-Latin-script target languages are in orange. Both metrics are correlated with their respective optimal convex combination of `src-en`, `en-tgt`, and `src-tgt` CLA scores, using the best-correlating sentence representation and CLA metric. Model: `Qwen3-Base`.

	Qwen3 base	Qwen3 inst.	gemma-3 base	gemma-3 inst.	Ministral-3 base	Ministral-3 inst.
$\rho_{\text{src-en}}$	81.8	81.7	66.7	77.8	**87.6**	63.4
repr.	fewshot	prompt	p-wm	fewshot	fewshot	last
metric	abs	ANC	ANC	ANC	abs	ANC
$\rho_{\text{src-tgt}}$	59.4	63.4	70.1	75.7	**81.1**	42.5
repr.	p-wm	prompt	p-wm	prompt	last	last
metric	ANC	ANC	ANC	ANC	abs	ANC
$\rho_{\oplus}$	82.0	81.7	75.8	82.9	**87.6**	63.4
impr.	0.2	0	4.6	**5.1**	0	0
$1-\alpha$	20	0	72	64	0	0
repr.	p-wm	prompt	p-wm	fewshot	fewshot	last
metric	ANC	ANC	ANC	ANC	abs	ANC
$\rho_{\oplus}^{\text{val}}$	75.2	72.2	67.2	75.1	**85.2**	53.6

Table 6: Best correlation between **Flores** translation CHRF and `src-en` (top), `src-tgt` (middle), and the optimal convex combination of `en-tgt` and `src-tgt` (bottom) CLA scores; **impr**ovement of the combination over `src-en` alone and the `src-tgt` weight ($1-\alpha$). The validation score ($\rho_{\oplus}^{\text{val}}$) applies the optimal parameters above to the **BOUQuET** dataset. Correlations are measured across `src` languages for a fixed `tgt`, averaged over target languages. Values in per cent.

main language set (44 languages)						
	Qwen3 base	Qwen3 inst.	gemma-3 base	gemma-3 inst.	Ministral-3 base	Ministral-3 inst.
$\rho_{\oplus}$	**91.2**	82.9	68.0	69.9	90.5	88.8
$1-\alpha-\beta$	0.0	8.0	0.0	**12.0**	4.0	0.0
repr.	p-wm	p-wm	last	p-wm	p-wm	p-wm
metric	cos	cos	ANC	cos	ratio	abs
$\rho_{\oplus}^{\text{val}}$	**91.1**	88.5	56.1	66.8	88.9	81.1
related language pairs (7 pairs)						
	Qwen3 base		gemma-3 base		Ministral-3 base	
$\rho_{\oplus}$	84.7		76.8		**90.0**	
$1-\alpha-\beta$	**88.0**		84.0		84.0	
repr.	mean		–		mean	
metric	cos		Eflo		cos	

Table 7: Best correlations between the optimal convex combination of `src-en`, `en-tgt`, and `src-tgt` and the **Flores** translation PMI for our main set of 44 languages (top) and a set of seven closely related languages (bottom; base models only). The validation score ($\rho_{\oplus}^{\text{val}}$) applies the optimal parameters above to the **BOUQuET** dataset. Values in per cent.

search for α and β as a heatmap. The results indicate moderate to strong correlations, also retained on the validation data. The **p-wm** sentence representation consistently performs the best, while **cos** appears to be the optimal metric for multiple models. Importantly, the weight of $\text{CLA}_{\texttt{src-tgt}}$ in the optimal convex combination is consistently near-zero or zero, suggesting that **most information about the translation quality is contained in the $\text{CLA}_{\texttt{src-en}}$ and the $\text{CLA}_{\texttt{en-tgt}}$ scores**.

Additionally, the results we show in Figure 5 also confirm our previous design decisions about PMI, showing that it can be used as a target-language independent metric (discussion in Appendix C).

6.4 What About Related Languages?

The findings from the previous section show that the alignment `src-tgt` has only little relevance for translation quality, implying that LLMs might internally translate first to English and then to the target language, as Wendler et al. (2024) suggested. However, in a practical setting, translating via English makes little sense when the source and target languages are closely related.

To examine this specific case, we create a new set of seven pairs of mutually intelligible languages, listed in Appendix A. Considering both directions, we work with 14 data points. We conduct the same experiments on these language pairs as in the pre-

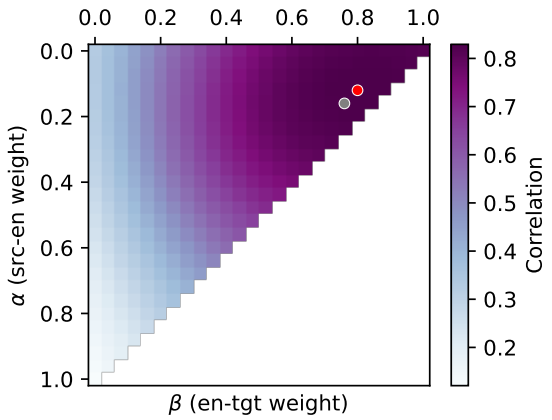

Figure 6: Correlation heatmap for the α and β grid search for the best-performing sentence representation/CLA metric pair for **`Qwen3 instr.`** when predicting **Flores** translation PMI. Values on the diagonal correspond to zero `src-tgt` weight. The best correlation is marked red. The best combination for **BOUQuET** is marked grey.

vious section. Table 7 (bottom) shows vastly different results than before: the weight for the `src-tgt` CLA score is now dominant ($\geq 84\%$), suggesting that the LLMs may not use English as a pivot point when direct translation is trivial.

7 CLA Metrics Evaluation

The results across all three tasks allow us to make several recommendations regarding CLA metric choice. While there is no one consistently best-performing approach to measuring cross-lingual alignment, the position-weighted mean (**p-wm**; Muennighoff, 2022) sentence representation wins in more settings than any other representation. Similarly, there was no single preferred alignment scoring method, but average neuron-wise correlation (**ANC**; Del and Fishel, 2022) wins in more settings than any other metric.

Therefore, we suggest using **p-wm** in combination with **ANC** in future experiments. However, in cases when sample-level alignment is needed, ANC is not applicable, and instead, we recommend using a retrieval-based metric (the simplest case is xSIM **abs**; Artetxe and Schwenk, 2019).

Compared to the embedding-level CLA scores, the tokeniser-level **Eflo**mal score (Hämmerl et al., 2025) achieves considerably worse correlation scores, except for the similar languages task. Nonetheless, in cases where the model's hidden states are unavailable (such as in closed-weight models), it can serve as a rough estimate of the model's expected performance.

8 Conclusions

We comprehensively review 27 approaches to measuring cross-lingual alignment from sentence embeddings by combining five sentence representation methods – one of which, **fewshot**, we newly propose – and six alignment scoring methods, together with one tokeniser-level metric. Our newly proposed method outperforms existing methods in numerous setups, showing potential for further development. We work with 44 languages of varying linguistic properties and three LLM families.

We assess how predictive CLA scores are for performance on three common NLP tasks: topic classification (SIB-200), reading comprehension (Belebele), and non-English-to-non-English machine translation. For monolingual tasks, we show that the performance in the `src` language is well correlated with the `src-en` CLA score, which is consistent with the findings from previous work (Kargaran et al., 2025; Ravisankar et al., 2026).

In the machine translation task, we address whether alignment between the source and target languages (`src-tgt`) is more predictive of translation quality than alignment via an intermediate English step (`src-en` and `en-tgt`). First, we measure translation quality using the CHRF score, which only allows us to analyse `src-tgt` and `src-en`. We find that `src-en` is generally more predictive of quality and only improves slightly by combining it with the `src-tgt` score.

Next, to include the `en-tgt` score in our analysis, we propose point-wise mutual information (PMI) for translation—a proxy for CHRF—which we show to be less dependent on the target language. Using this method, we show that the `src-tgt` score does not further improve the predictions of the `src-en` and `en-tgt` combined. This contributes evidence to the claim that LLMs use English as a pivot language (Wendler et al., 2024) from a yet unstudied perspective.

We summarise the predictive performance of different CLA scores into three recommendations for best performance estimates: (1) Use **p-wm+ANC** for system-level tasks scoring. (2) Use **p-wm+abs** when sample-level scores are needed. (3) Use **Eflo**mal only when model weights are unavailable.

Limitations

Domain coverage. Highly multilingual parallel data remains scarce, and the Flores-200 dataset, while the best available resource for our purposes, covers a single domain. Consequently, our CLA scores and two downstream tasks share the same domain, which might inflate the observed correlations compared to what one would expect in more diverse settings. Controlling for domain effects would require multilingual parallel corpora of comparable quality across multiple domains. Such a dataset (BOUQuET) exists for translation, which we use for validation. However, to our knowledge, there is no suitable alternative to the SIB-200 and Belebele dataset that covers a different domain with the same level of multilinguality.

The few-shot sentence representation. Our proposed **fewshot** embedding method relies on Wikidata entity descriptions, which vary in quality and may not describe the same entities across all 44 languages. A more principled approach would use verified multilingual dictionary definitions, but no such resource exists with the language coverage we require. Nonetheless, our proposed method remained the best-performing for certain setups, highlighting its potential.

Target-language independence of PMI. While we show empirically that our proposed PMI metric clusters target languages less severely than CHRF, full independence cannot be formally established. The metric remains sensitive to model-specific calibration, and the degree of residual dependence likely varies across models. We view this as an invitation for future work rather than a deficiency of the current proposal.

Related languages experiment. Our side experiment with related languages (Section 6.4) covers only 7 pairs of related languages and is evaluated only on base models. Furthermore, token overlap between related languages may skew the PMI and CHRF distributions. We invite future work to examine the role of English as a pivot language in more detail when translating between related languages.

Language coverage. We selected the set of 44 languages we work with to cover various families, resource-richnesses, and writing systems. 17 of those languages (37%) are European, which is not representative of the global population. Nonetheless, our experiments (Figure 5) show that the script and the resource-richness of the language affect the results more than the language's geographical origin.

Ethical Considerations

We do not anticipate any negative ethical implications arising from this study. We used AI-assisted coding (GitHub Copilot) and AI tools (Claude, ChatGPT, and Grammarly) for grammar checking and minor sentence rephrasing.

The inspected open-weight models are each licensed under the `Apache 2.0` license (`Qwen3` and `Ministral-3`) and the `gemma` license (`gemma-3`). The BOUQuET dataset is licensed under `CC-BY-4.0`, while the remaining datasets we work with are licensed under `CC-BY-SA-4.0`.

Acknowledgments

Adnan was supported by the HumanAId project `CZ.02.01.01/00/23_025/0008691` of the Czech Ministry of Education. Jindřich was supported by the CUNI project `PRIMUS/23/SCI/023` and project `CZ.02.01.01/00/23_020/0008518` of the Czech Ministry of Education. This research was partially supported by SVV project number `260 821`.

This work was co-funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie 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.

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET : dataset, benchmark and open initiative for universal quality evaluation

in translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

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 and virtual meeting. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Maksym Del and Mark Fishel. 2022. Cross-lingual similarity of multilingual representations revisited. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 185–195, Online only. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Katharina Hämmerl, Tomasz Limisiewicz, Jindřich Libovický, and Alexander Fraser. 2025. Beyond literal token overlap: Token alignability for multilinguality. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 756–767, Albuquerque, New Mexico. Association for Computational Linguistics.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schuetze. 2025. MEXA: Multilingual evaluation of English-centric LLMs via cross-lingual alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 27001–27023, Vienna, Austria. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *Preprint*, arXiv:2202.08904.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Zhi Qu, Chenchen Ding, and Taro Watanabe. 2025. Languages transferred within the encoder: On representation transfer in zero-shot multilingual translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 81–98, Geneva, Switzerland. European Association for Machine Translation.

Kartik Ravisankar, Hyojung Han, Sarah Wiegreffe, and Marine Carpuat. 2026. Can you map it to english? the role of cross-lingual alignment in multilingual performance of llms. *Preprint*, arXiv:2504.09378.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2021. DefSent: Sentence embeddings using definition sentences. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 411–418, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Evaluated Languages

In this article, we work with the following languages: Afrikaans (`af`), Amharic (`am`), Arabic (`ar`), Azerbaijani (`az`), Bulgarian (`bg`), Bengali (`bn`), Czech (`cs`), Danish (`da`), German (`de`), Greek (`el`), English (`en`), Spanish (`es`), Farsi (`fa`), Finnish (`fi`), French (`fr`), Hebrew (`he`), Hindi (`hi`), Hungarian (`hu`), Armenian (`hy`), Indonesian (`id`), Icelandic (`is`), Italian (`it`), Japanese (`ja`), Javanese (`jv`), Georgian (`ka`), Central Khmer (`km`), Kannada (`kn`), Korean (`ko`), Lithuanian (`lt`), Malayalam (`ml`), Marathi (`mr`), Dutch (`nl`), Polish (`pl`), Portuguese (`pt`), Russian (`ru`), Swahili (`sw`), Tamil (`ta`), Telugu (`te`), Thai (`th`), Tagalog (`tl`), Turkish (`tr`), Urdu (`ur`), Vietnamese (`vi`), and Chinese (`zh`).

In Section 6.4 only, we work with a special set of seven pairs of closely related languages. Note that we use each pair in both directions, yielding 14 data points: Czech–Slovak (`cs-sk`), Spanish–Catalan (`es-ca`), Afrikaans–Dutch (`af-nl`), Latvian–Lithuanian (`lv-lt`), Turkish–Azerbaijani (`tr-az`), Danish–Norwegian (`da-nb`), and Urdu–Hindi (`ur-hi`).

B Inference and Prompts

For high-level inference, we use the vLLM inference engine (Kwon et al., 2023). Elsewhere, we use Transformers (Wolf et al., 2020) with code optimisation for key-value caching. We run the experiments on 1–2 NVIDIA A40 (48 GB) or A100 (40 GB) GPUs, depending on the task.

The following are example prompts for each task. The empty red circle symbol (○) denotes an empty line. The curled arrow (↪) denotes text wrapping in the box, i.e. there is no newline character in the place of the arrow.

```
totality consisting of space, time,
 ↪ matter and energy
universe
○
third planet from the Sun in the Solar
 ↪ System
Earth
○
matter capable of extracting energy
 ↪ from the environment for replication
life
○
permanent cessation of vital functions
death
○
any single member of Homo sapiens,
 ↪ unique extant species of the genus
 ↪ Homo
human
○
All human beings are born small.
□
```

Listing 1: Example **few-shot** prompt for embedding the English sentence ‘All human beings are born small.’ The embeddings would be extracted from the last token, denoted by □ (the symbol is not a part of the actual prompt).

```
Classify the following text into one of
 ↪ these topics: "science/technology",
 ↪ "travel", "politics", "sports",
 ↪ "health", "entertainment",
 ↪ "geography". Provide only the topic
 ↪ in English as your response.
○
text: `<text segment>`
topic: `
```

Listing 2: Prompt for extracting the *SIB-200* label. The English prompt is used for all text segments regardless of the language.

```
Make sure your hand is as relaxed as
 ↪ possible while still hitting all
 ↪ the~notes correctly - also try not
 ↪ to make much extraneous motion with
 ↪ your fingers. This way, you will
 ↪ tire yourself out as little as
 ↪ possible. Remember there's no need
 ↪ to hit the keys with a lot of force
 ↪ for extra volume like on the piano.
 ↪ On the accordion, to get extra
 ↪ volume, you use the bellows with
 ↪ more pressure or speed.
○
According to the passage, what would
 ↪ not be considered an accurate tip
 ↪ for successfully playing the
 ↪ accordion?
○
For additional volume, increase the
 ↪ force with which you hit the keys
Keep unnecessary movement to a minimum
 ↪ in order to preserve your stamina
Be mindful of hitting the notes while
 ↪ maintaining a relaxed hand
Increase the speed with which you
 ↪ operate the bellows to achieve
 ↪ extra volume
○
For additional volume, increase the
 ↪ force with which you hit the keys
○
○
One of the most common problems when
 ↪ trying to convert a movie to DVD
 ↪ format is the overscan. Most
 ↪ televisions are made in a way to
 ↪ please the general public. For that
 ↪ reason, everything you see on the
 ↪ TV had the borders cut, top, bottom
 ↪ and sides. This is made to ensure
 ↪ that the image covers the whole
 ↪ screen. That is called overscan.
 ↪ Unfortunately, when you make a DVD,
 ↪ it's borders will most likely be
 ↪ cut too, and if the video had
 ↪ subtitles too close to the bottom,
 ↪ they won't be fully shown.
○
Why do the images on television have
 ↪ their borders cut?
○
To allow for subtitles
So the image fills the entire screen
To allow for simple conversion into
 ↪ other formats
To cut subtitles too close to the bottom
```

Listing 3: Example prompt for a *Belebele* sample in English, using one reference shot (we use three in the experiments). Supposing that each word is tokenised into a single token, the model would only need to generate 4 tokens, which suffices to determine the answer uniquely ('To allow for subtitles' vs. 'To allow for simple [...]').

```
USER: Make sure your hand is as relaxed
 ↪ as possible while still hitting all
 ↪ the~notes correctly - also try not
 ↪ to make much extraneous motion with
 ↪ your fingers. This way, you will
 ↪ tire yourself out as little as
 ↪ possible. Remember there's no need
 ↪ to hit the keys with a lot of force
 ↪ for extra volume like on the piano.
 ↪ On the accordion, to get extra
 ↪ volume, you use the bellows with
 ↪ more pressure or speed.
○
According to the passage, what would
 ↪ not be considered an accurate tip
 ↪ for successfully playing the
 ↪ accordion?
○
For additional volume, increase the
 ↪ force with which you hit the keys
Keep unnecessary movement to a minimum
 ↪ in order to preserve your stamina
Be mindful of hitting the notes while
 ↪ maintaining a relaxed hand
Increase the speed with which you
 ↪ operate the bellows to achieve
 ↪ extra volume
ASSISTANT: For additional volume,
 ↪ increase the force with which you
 ↪ hit the keys
USER: One of the most common problems
 ↪ when trying to convert a movie to
 ↪ DVD format is the overscan. Most
 ↪ televisions are made in a way to
 ↪ please the general public. For that
 ↪ reason, everything you see on the
 ↪ TV had the borders cut, top, bottom
 ↪ and sides. This is made to ensure
 ↪ that the image covers the whole
 ↪ screen. That is called overscan.
 ↪ Unfortunately, when you make a DVD,
 ↪ it's borders will most likely be
 ↪ cut too, and if the video had
 ↪ subtitles too close to the bottom,
 ↪ they won't be fully shown.
○
Why do the images on television have
 ↪ their borders cut?
○
To allow for subtitles
So the image fills the entire screen
To allow for simple conversion into
 ↪ other formats
To cut subtitles too close to the bottom
```

Listing 4: Example prompt for a *Belebele* sample in English for instruction-tuned models, using one reference shot (we use three in the experiments). Supposing that each word is tokenised into a single token, the model would only need to generate 4 tokens, which suffices to determine the answer uniquely ('To allow for subtitles' vs. 'To allow for simple [...]'). We further use guided decoding that disables generating the * markdown formatting symbol.

```
"We now have 4-month-old mice that are
 ↪ non-diabetic that used to be
 ↪ diabetic," he added.
"Abbiamo topi di quattro mesi che prima
 ↪ erano diabetici e ora non lo sono
 ↪ più", ha aggiunto.
○
Dr. Ehud Ur, professor of medicine at
 ↪ Dalhousie University in Halifax,
 ↪ Nova Scotia and chair of the
 ↪ clinical and scientific division of
 ↪ the Canadian Diabetes Association
 ↪ cautioned that the research is
 ↪ still in its early days.
Lo studio è ancora in fase iniziale,
 ↪ come dichiarato cautelativamente
 ↪ dal dottor Ehud Ur, professore di
 ↪ medicina alla Dalhousie University
 ↪ di Halifax, Nuova Scozia, e
 ↪ direttore del dipartimento clinico
 ↪ e scientifico della Canadian
 ↪ Diabetes Association.
○
Like some other experts, he is
 ↪ skeptical about whether diabetes
 ↪ can be cured, noting that these
 ↪ findings have no relevance to
 ↪ people who already have Type 1
 ↪ diabetes.
Come altri esperti, è scettico circa la
 ↪ possibilità di curare il diabete,
 ↪ sottolineando che questi risultati
 ↪ non hanno rilevanza per chi è già
 ↪ affetto dal diabete di tipo 1.
○
Previously, Ring's CEO, Jamie Siminoff,
 ↪ remarked the company started when
 ↪ his doorbell wasn't audible from
 ↪ his shop in his garage.
```

Listing 5: Example prompt for translating the English sentence '*Previously, Ring's CEO, Jamie Siminoff, remarked the company started when his doorbell wasn't audible from his shop in his garage.*' to Italian.

```
USER: Classify the following text into
 ↪ one of these topics:
 ↪ "science/technology", "travel",
 ↪ "politics", "sports", "health",
 ↪ "entertainment", "geography".
 ↪ Provide only the topic in English
 ↪ as your response.
○
text: `<text segment>`
 ASSISTANT: The topic is: `
```

Listing 6: Prompt for extracting the *SIB-200* label for instruction-tuned LLMs. The English prompt is used for all text segments regardless of the language.

```
"Abbiamo topi di quattro mesi che prima
 ↪ erano diabetici e ora non lo sono
 ↪ più", ha aggiunto.
Lo studio è ancora in fase iniziale,
 ↪ come dichiarato cautelativamente
 ↪ dal dottor Ehud Ur, professore di
 ↪ medicina alla Dalhousie University
 ↪ di Halifax, Nuova Scozia, e
 ↪ direttore del dipartimento clinico
 ↪ e scientifico della Canadian
 ↪ Diabetes Association.
Come altri esperti, è scettico circa la
 ↪ possibilità di curare il diabete,
 ↪ sottolineando che questi risultati
 ↪ non hanno rilevanza per chi è già
 ↪ affetto dal diabete di tipo 1.
I media locali riferiscono che durante
 ↪ l'intervento un veicolo
 ↪ aeroportuale antincendio si è
 ↪ ribaltato.
```

Listing 7: Example prompt for extracting the prior probability of the Italian sentence '*I media locali riferiscono che durante l'intervento un veicolo aeroportuale antincendio si è ribaltato.*'.

C Target-Language Independence of PMI

While the goal of the PMI translation metric is target-language independence, it is difficult to validate this property, as LLMs a priori perform better in some languages, making direct comparisons difficult. Nonetheless, we propose that a strong correlation within mixed target languages (such as that shown in Table 7, top; for experimental setup see Section 6.3) indicates target-language independence. Note that this is only a one-sided implication, as poor correlation can be attributed to various factors.

Figure 5 shows the comparison in correlation between PMI (left) and CHRF (right). The plot for CHRF score – which is target-language-sensitive – contains clusters of target languages which do not follow the general pattern: most notably Chinese, for which the CHRF score is considerably smaller than a linear classifier would predict from the CLA scores. Such inconsistencies reduce correlation. While outlying target-language clusters are also present in the PMI plot, they are less pronounced, resulting in higher correlation scores. Furthermore, as Figures 7–9 show, the clusters are not consistent across models and seem to be nuances of the models rather than the metric itself. Therefore, while we could not validate that our proposed PMI metric is target-language-independent, empirically, it is *less* dependent than CHRF and, unlike CHRF, does not exhibit systematic biases against specific languages across different models.

D Detailed Results

Qwen3-14B-Base	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	26.9	29.6	61.0	61.0	61.3	–	16.8
last	25.6	29.5	61.8	61.8	62.0	18.7	
p-wm	40.5	54.0	79.4	79.4	79.5	–	
mean	21.4	54.7	64.8	64.8	64.9	–	
prompt	–	59.4	**81.8**	**81.8**	**81.8**	–	

Qwen3-14B	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	59.7	44.1	44.1	44.1	–	16.6
last	16.1	28.2	63.8	63.8	63.7	53.5	
p-wm	38.4	65.5	78.5	78.5	78.5	–	
mean	18.9	67.2	75.7	75.7	75.9	–	
prompt	–	81.6	**85.9**	**85.9**	**85.9**	–	

gemma-3-12b-pt	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	58.2	62.1	70.0	70.0	70.0	–	49.9
last	20.3	65.4	65.7	65.7	65.7	62.4	
p-wm	**70.9**	62.8	58.8	58.8	58.8	–	
mean	–	61.9	62.4	62.4	62.4	–	
prompt	–	33.3	55.5	55.5	55.3	–	

gemma-3-12b-it	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	12.1	40.3	53.2	53.2	53.1	–	24.5
last	48.7	45.7	50.6	50.6	50.5	9.6	
p-wm	60.9	49.6	61.0	61.0	61.0	–	
mean	–	49.4	**67.6**	**67.6**	**67.6**	–	
prompt	–	50.9	62.0	62.0	62.0	–	

Ministral-3-14B-Base-2512	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	56.4	74.7	89.5	89.5	89.5	–	45.2
last	73.0	66.1	73.4	73.4	73.2	92.9	
p-wm	91.9	87.2	83.8	83.8	84.1	–	
mean	69.4	87.0	80.3	80.3	80.8	–	
prompt	–	92.5	**97.5**	**97.5**	**97.5**	–	

Ministral-3-14B-Instruct-2512	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	86.7	95.2	95.2	95.2	–	36.7
last	70.8	64.4	92.3	92.3	92.3	88.3	
p-wm	89.5	84.7	80.9	80.9	80.9	–	
mean	65.4	80.3	71.7	71.7	72.3	–	
prompt	9.7	87.5	**97.4**	**97.4**	**97.4**	–	

Table 8: Correlation measured across the `src` languages of `src-en` alignment score and the **SIB-200** F_1 score in the `src` language for all models. Valuesin cent.

Qwen3-14B-Base	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	47.8	50.5	91.4	91.4	91.5	–	50.2
last	42.2	51.2	81.4	81.4	81.4	24.3	
p-wm	80.3	91.0	89.7	89.7	89.7	–	
mean	59.1	91.4	89.3	89.3	89.3	–	
prompt	–	**91.7**	85.4	85.4	85.3	–	

Qwen3-14B	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	89.6	83.3	83.3	83.3	–	51.8
last	42.3	56.0	81.7	81.7	81.6	80.8	
p-wm	76.4	91.8	90.6	90.6	90.6	–	
mean	58.0	92.4	86.4	86.4	86.4	–	
prompt	–	**95.2**	81.9	81.9	81.9	–	

gemma-3-12b-pt	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	46.5	59.3	56.4	56.4	56.3	–	24.2
last	32.3	49.4	59.9	59.9	59.9	42.6	
p-wm	62.5	66.6	**70.0**	**70.0**	**70.0**	–	
mean	–	66.7	64.0	64.0	64.0	–	
prompt	–	31.1	59.6	59.6	59.2	–	

gemma-3-12b-it	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	20.3	81.8	82.6	82.6	82.6	–	58.9
last	68.5	69.2	80.6	80.6	80.7	12.6	
p-wm	81.7	**85.0**	69.2	69.2	69.2	–	
mean	–	84.4	65.2	65.2	65.2	–	
prompt	–	83.7	65.9	65.9	65.9	–	

Ministral-3-14B-Base-2512	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	58.8	78.0	95.4	95.4	95.4	–	53.3
last	88.0	84.8	89.3	89.3	89.3	89.6	
p-wm	90.3	**97.4**	95.5	95.5	95.7	–	
mean	74.0	**97.4**	93.7	93.7	93.9	–	
prompt	–	93.5	87.8	87.8	87.8	–	

Ministral-3-14B-Instruct-2512	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	94.6	**95.7**	**95.7**	**95.7**	–	50.8
last	72.8	73.6	94.3	94.3	94.4	93.1	
p-wm	92.5	94.2	91.8	91.8	91.8	–	
mean	74.2	91.8	84.8	84.8	85.1	–	
prompt	4.2	**95.6**	94.3	94.3	94.3	–	

Table 9: Correlation measured across the `src` languages of `src-en` alignment score and the **Belebele** accuracy in the `src` language for all models. Values in per cent.

	Qwen3		gemma-3		Ministral-3	
	base	inst.	base	inst.	base	inst.
af	45.6	45.3	**51.4**	49.3	46.7	29.6
am	7.9	8.8	**24.5**	20.6	4.6	6.5
ar	39.1	38.0	44.2	**44.7**	40.3	21.7
az	29.1	28.1	**37.2**	36.8	32.6	24.5
bg	44.8	40.4	**52.6**	**52.6**	46.8	15.1
bn	29.8	26.4	**39.7**	38.8	32.7	18.4
cs	40.2	39.9	**46.1**	**46.2**	42.7	23.7
da	46.5	46.6	**54.3**	54.0	48.2	22.8
de	47.6	47.7	51.0	**51.9**	49.4	23.7
el	36.7	37.2	**44.0**	43.8	39.6	21.8
en	59.4	58.4	**62.4**	60.0	59.4	45.6
es	45.3	45.5	47.5	**47.8**	45.7	24.4
fa	38.3	37.5	45.0	**45.6**	41.9	23.7
fi	39.2	38.8	46.6	**47.1**	42.7	29.1
fr	52.2	52.0	55.7	**56.0**	52.4	24.9
he	33.1	30.9	**45.7**	45.3	39.9	17.4
hi	36.2	34.4	44.5	**44.7**	41.2	12.7
hu	39.8	40.1	**44.6**	**44.6**	40.6	26.3
hy	33.8	26.9	**43.4**	40.4	39.5	25.4
id	52.8	52.4	56.6	**57.3**	50.1	25.4
is	27.9	27.5	**39.8**	36.5	34.0	23.7
it	46.0	46.0	49.7	**50.1**	47.1	18.7
ja	23.8	23.9	27.0	**27.3**	25.2	14.3
jv	35.8	35.6	**43.6**	41.8	25.7	23.7
ka	31.5	29.9	**40.4**	38.8	35.7	26.5
km	21.7	18.3	**28.4**	26.4	9.1	12.4
kn	27.3	23.1	**41.4**	39.2	34.8	19.8
ko	25.4	22.9	29.0	**29.8**	25.6	15.0
lt	38.6	37.5	**46.3**	44.9	39.8	20.9
ml	26.2	24.6	**41.5**	36.6	31.5	21.0
mr	26.8	27.4	38.3	**38.6**	33.2	10.7
nl	44.3	44.5	**48.4**	**48.5**	43.8	18.9
pl	38.4	38.8	42.8	**43.5**	40.3	26.0
pt	51.7	51.0	**55.4**	55.3	52.3	29.6
ru	43.8	42.9	47.5	**47.8**	45.2	19.9
sw	28.9	28.0	**49.0**	47.0	34.9	26.6
ta	30.6	32.9	42.2	**44.9**	35.7	22.3
te	29.8	30.4	40.0	**43.0**	31.6	15.3
th	35.4	35.2	39.7	**40.2**	35.2	25.2
tl	45.2	44.0	52.0	**52.3**	45.0	23.9
tr	41.7	39.0	47.7	**47.8**	43.3	23.6
ur	27.7	28.7	**39.2**	38.2	34.6	17.9
vi	48.2	46.2	**51.6**	**51.6**	47.3	24.5
zh	19.5	19.3	**21.5**	**21.4**	20.5	8.7

Table 10: Average **Flores** translation performance (CHRF; over all source languages) for the displayed target languages. Values in cent.

	Qwen3		gemma-3		Ministral-3	
	base	inst.	base	inst.	base	inst.
af	46.3	43.8	**53.7**	49.9	45.6	21.5
am	10.6	9.8	**26.3**	20.5	6.3	5.8
ar	27.1	25.8	**36.3**	35.4	28.7	13.5
az	30.7	27.1	**42.8**	40.1	33.6	19.8
bg	47.5	41.3	**56.1**	52.7	48.4	14.0
bn	38.9	30.0	**48.0**	39.8	37.0	9.0
cs	43.0	39.6	**50.6**	47.9	44.9	28.2
da	46.3	43.2	**53.1**	50.2	47.6	17.4
de	49.9	47.7	**53.8**	51.2	50.8	17.6
el	42.8	39.3	**52.8**	48.1	46.1	17.8
en	60.1	56.0	**63.0**	56.9	59.2	25.3
es	52.3	49.2	**55.4**	52.5	52.0	26.6
fa	35.3	32.7	**42.4**	40.4	37.3	15.0
fi	39.1	33.4	**47.6**	46.1	41.0	20.5
fr	52.6	48.9	**56.1**	51.5	53.2	27.0
he	36.8	30.3	**47.1**	45.4	38.6	17.3
hi	38.0	32.9	**43.6**	38.1	39.9	8.0
hu	44.1	40.3	**49.6**	46.5	43.6	20.0
hy	35.8	19.0	**44.7**	38.8	36.3	21.3
id	50.9	46.1	**53.9**	51.1	48.5	13.6
is	30.5	27.9	**44.2**	40.3	34.2	17.4
it	54.4	51.1	**59.0**	56.2	55.3	23.9
ja	23.3	20.8	**25.8**	24.5	24.0	12.8
jv	36.4	32.6	**47.1**	41.2	26.5	18.4
ka	32.8	28.3	**42.5**	37.5	32.9	23.7
km	31.9	23.0	**38.7**	32.3	14.6	12.6
kn	32.4	22.5	**41.5**	32.5	34.5	9.1
ko	28.5	23.9	**33.0**	31.6	28.9	12.7
lt	42.6	39.6	**51.7**	47.5	42.7	25.8
ml	31.2	25.3	**38.4**	34.0	30.2	14.0
mr	31.5	26.6	**38.5**	36.5	33.3	7.9
nl	53.0	48.5	**59.0**	54.9	53.6	24.1
pl	47.9	42.9	**55.4**	51.5	50.3	21.5
pt	55.7	52.7	**60.4**	56.6	56.6	27.6
ru	47.9	42.0	**52.2**	48.1	49.2	12.6
sw	27.7	23.2	**50.5**	46.6	31.5	19.8
ta	35.4	27.6	**42.5**	41.2	36.8	11.9
te	34.2	28.8	**42.4**	41.2	34.5	10.6
th	38.8	34.0	**43.3**	41.2	36.4	19.8
tl	47.1	40.2	**56.8**	50.1	43.4	17.9
tr	45.8	38.7	**53.0**	50.2	45.9	19.9
ur	37.8	32.8	**49.6**	45.3	39.5	14.1
vi	55.7	43.0	**59.8**	54.4	53.2	25.2
zh	24.5	20.2	**26.6**	23.4	25.1	10.1

Table 11: Average **BOUQuET** translation performance (CHRF; over all source languages) for the displayed target languages. Values in cent.

Qwen3-14B-Base

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	43/20	56/48	**82**/56	**82**/56	**82**/56	–/–	62/15
last	34/16	57/48	72/52	72/52	72/53	17/14	
p-wm	77/36	**82/59**	72/**59**	72/**59**	72/**59**	–/–	
mean	61/30	81/57	77/28	77/28	76/28	–/–	
prompt	–/–	80/58	67/58	67/58	67/58	–/–	

Qwen3-14B

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–/–	80/50	71/43	71/43	71/43	–/–	50/7
last	34/18	54/44	69/45	69/45	69/45	69/58	
p-wm	68/30	80/58	74/58	74/58	74/58	–/–	
mean	52/29	81/57	73/27	73/27	73/27	–/–	
prompt	–/–	**82/63**	63/55	63/55	63/55	–/–	

gemma-3-12b-pt

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	43/28	57/42	51/22	51/22	51/22	–/–	65/53
last	22/10	57/50	64/58	64/58	64/58	53/50	
p-wm	57/53	**67/70**	54/52	54/52	54/52	–/–	
mean	–/20	**67**/68	49/46	49/46	49/46	–/–	
prompt	–/–	60/66	50/54	50/54	51/54	–/–	

gemma-3-12b-it

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	20/25	**78**/70	71/53	71/53	71/53	–/–	68/48
last	52/44	71/51	71/54	71/54	71/54	26/26	
p-wm	66/53	77/70	58/48	58/48	58/48	–/–	
mean	–/27	77/68	55/43	55/43	55/42	–/–	
prompt	–/–	73/**76**	62/63	62/63	62/63	–/–	

Ministral-3-14B-Base-2512

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	56/43	77/74	**88**/77	**88**/77	**88**/77	–/–	63/30
last	82/64	79/70	83/**81**	83/**81**	83/**81**	85/78	
p-wm	84/72	87/77	86/75	86/75	86/75	–/–	
mean	67/60	87/77	84/64	84/64	84/64	–/–	
prompt	–/–	87/78	81/**81**	81/**81**	81/**81**	–/–	

Ministral-3-14B-Instruct-2512

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–/–	56/32	35/28	35/28	35/28	–/–	62/29
last	46/33	**63/43**	42/36	42/36	42/36	43/33	
p-wm	39/23	52/28	54/27	54/27	54/27	–/–	
mean	27/21	55/29	60/23	60/23	59/22	–/–	
prompt	2/-15	50/31	28/24	28/24	28/24	–/–	

Table 12: Correlation between the CLA and the **Flores** translation CHRF measured first across the `src` languages with a fixed `tgt` language and then averaged over the target languages. The correlation is measured between the `src-en` (left of the '/' symbol) or `src-tgt` (right of the '/' symbol) alignment score and the CHRF score in the `tgt` language. Values in per cent. Cell colours are based on the `src-en` scores.

Qwen3-14B-Base

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	43/0	63/**7**	**82**/0	**82**/0	**82**/0	–/–	62/0
last	34/0	59/2	72/0	72/0	72/0	17/0	
p-wm	79/2	**82**/0	72/0	72/0	72/0	–/–	
mean	61/0	**82**/0	77/0	77/0	76/0	–/–	
prompt	–/–	80/0	67/0	67/0	67/0	–/–	

Qwen3-14B

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–/–	80/0	73/2	73/2	73/2	–/–	50/0
last	34/0	56/2	69/0	69/0	69/0	69/1	
p-wm	71/**3**	81/0	74/0	74/0	74/0	–/–	
mean	52/0	81/0	73/0	73/0	73/0	–/–	
prompt	–/–	**82**/0	63/0	63/0	63/0	–/–	

gemma-3-12b-pt

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	45/2	59/3	51/0	51/0	51/0	–/–	69/4
last	22/0	62/5	72/7	72/7	72/7	58/5	
p-wm	65/8	**76**/9	57/3	57/3	57/3	–/–	
mean	20/–	75/8	52/3	52/3	52/3	–/–	
prompt	–/–	74/**14**	57/7	57/7	58/7	–/–	

gemma-3-12b-it

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	27/**7**	**83**/5	72/1	72/1	72/1	–/–	70/2
last	54/2	72/1	76/4	76/4	75/4	26/0	
p-wm	71/5	80/3	59/1	59/1	59/1	–/–	
mean	27/–	80/3	55/0	55/0	55/0	–/–	
prompt	–/–	79/6	65/2	65/2	65/2	–/–	

Ministral-3-14B-Base-2512

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	56/0	80/**3**	**88**/0	**88**/0	**88**/0	–/–	63/0
last	82/0	81/1	84/1	84/1	84/1	85/0	
p-wm	84/0	87/0	86/0	86/0	86/0	–/–	
mean	67/0	87/0	85/0	85/0	85/0	–/–	
prompt	–/–	87/0	82/0	82/0	82/0	–/–	

Ministral-3-14B-Instruct-2512

	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–/–	56/**0**	35/**0**	35/**0**	35/**0**	–/–	62/**0**
last	46/**0**	**63/0**	42/**0**	42/**0**	42/**0**	43/**0**	
p-wm	39/**0**	52/**0**	54/**0**	54/**0**	54/**0**	–/–	
mean	27/**0**	55/**0**	60/**0**	60/**0**	59/**0**	–/–	
prompt	2/**0**	50/**0**	28/**0**	28/**0**	28/**0**	–/–	

Table 13: Correlation between the CLA and the **Flores** translation CHRF measured first across the `src` languages with a fixed `tgt` language and then averaged over the target languages. The correlation is measured between the optimal convex combination of `src-en` and `src-tgt` alignment scores and the CHRF score in the `tgt` language. Values are displayed as per cent. The improvement of the combination over the `en-tgt` (within the same metric-representation) is shown right to the '/' symbol.

Qwen3-14B-Base							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	60.7	63.3	80.1	80.1	80.1	–	84.3
last	50.2	62.3	54.2	54.2	54.4	28.1	
p-wm	**91.2**	81.1	57.5	57.5	57.5	–	
mean	85.2	79.5	90.8	90.8	90.7	–	
prompt	–	71.8	43.0	43.0	42.9	–	

Qwen3-14B							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	74.2	79.3	79.3	79.3	–	74.9
last	44.8	55.6	52.0	52.0	52.0	66.5	
p-wm	**82.9**	75.1	55.6	55.6	55.6	–	
mean	74.8	74.1	79.9	79.9	79.9	–	
prompt	–	64.0	36.2	36.2	36.3	–	

gemma-3-12b-pt							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	30.2	46.0	48.2	48.2	48.2	–	54.5
last	15.8	**68.0**	67.3	67.3	65.7	37.8	
p-wm	64.8	42.0	53.7	53.7	53.7	–	
mean	2.9	41.8	56.0	56.0	56.0	–	
prompt	–	39.4	47.5	47.5	47.4	–	

gemma-3-12b-it							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	26.4	50.7	60.7	60.7	60.7	–	48.0
last	55.6	59.0	67.2	67.2	67.7	32.9	
p-wm	**69.9**	48.8	62.2	62.2	62.2	–	
mean	8.4	48.1	63.7	63.7	63.7	–	
prompt	–	50.8	38.9	38.9	38.9	–	

Ministral-3-14B-Base-2512							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	57.4	79.6	85.7	85.7	85.8	–	69.7
last	84.6	83.9	87.0	87.0	87.1	80.8	
p-wm	81.3	84.2	**90.4**	**90.4**	**90.5**	–	
mean	63.3	83.9	88.8	88.8	88.9	–	
prompt	–	79.6	73.2	73.2	73.1	–	

Ministral-3-14B-Instruct-2512							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	84.2	76.5	76.5	76.5	–	69.4
last	67.0	78.4	75.1	75.1	75.1	79.1	
p-wm	79.3	82.2	**88.8**	**88.8**	**88.7**	–	
mean	60.4	82.1	85.2	85.2	85.2	–	
prompt	19.3	80.8	70.5	70.5	70.5	–	

Table 14: Correlation between the optimal convex combination of alignments `src-en`, `en-tgt` and `src-tgt` and the **Flores** translation PMI. Values in per cent.

Qwen3-14B-Base							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	0.0	**40.0**	0.0	0.0	0.0	–	0.0
last	0.0	20.0	0.0	0.0	0.0	0.0	
p-wm	0.0	0.0	0.0	0.0	0.0	–	
mean	0.0	0.0	24.0	24.0	24.0	–	
prompt	–	0.0	0.0	0.0	0.0	–	

Qwen3-14B							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	0.0	8.0	8.0	8.0	–	0.0
last	0.0	16.0	0.0	0.0	0.0	0.0	
p-wm	8.0	0.0	0.0	0.0	0.0	–	
mean	0.0	0.0	**36.0**	**36.0**	**36.0**	–	
prompt	–	0.0	0.0	0.0	0.0	–	

gemma-3-12b-pt							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	0.0	0.0	0.0	0.0	0.0	–	52.0
last	0.0	0.0	0.0	0.0	0.0	24.0	
p-wm	0.0	0.0	0.0	0.0	0.0	–	
mean	**72.0**	0.0	0.0	0.0	0.0	–	
prompt	–	16.0	0.0	0.0	0.0	–	

gemma-3-12b-it							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	0.0	12.0	0.0	0.0	0.0	–	36.0
last	12.0	0.0	0.0	0.0	0.0	0.0	
p-wm	0.0	12.0	0.0	0.0	0.0	–	
mean	**72.0**	16.0	0.0	0.0	0.0	–	
prompt	–	0.0	0.0	0.0	0.0	–	

Ministral-3-14B-Base-2512							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	0.0	40.0	0.0	0.0	0.0	–	4.0
last	4.0	32.0	32.0	32.0	32.0	0.0	
p-wm	0.0	0.0	4.0	4.0	4.0	–	
mean	4.0	0.0	24.0	24.0	24.0	–	
prompt	–	4.0	56.0	56.0	**64.0**	–	

Ministral-3-14B-Instruct-2512							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	0.0	0.0	0.0	0.0	–	0.0
last	4.0	24.0	0.0	0.0	0.0	0.0	
p-wm	0.0	0.0	0.0	0.0	0.0	–	
mean	0.0	0.0	24.0	24.0	24.0	–	
prompt	0.0	0.0	**32.0**	**32.0**	**32.0**	–	

Table 15: Optimal convex-combination weight for the `src-tgt` alignment ($1-\alpha-\beta$) for predicting the **Flores** translation PMI scores for all models. Values are shown as per cent.

Qwen3-14B-Base							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	33/0	61/**13**	73/1	73/1	73/0	–/–	51/0
last	23/0	55/6	69/0	69/0	69/0	23/0	
p-wm	70/4	75/2	68/0	68/0	68/0	–/–	
mean	50/1	74/1	70/0	70/0	70/0	–/–	
prompt	–/–	**76**/0	65/0	65/0	65/0	–/–	
Qwen3-14B							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–/–	69/1	61/2	61/2	61/2	–/–	37/0
last	23/0	48/**3**	61/0	61/0	61/0	59/1	
p-wm	57/**3**	70/1	65/0	65/0	65/0	–/–	
mean	39/0	70/1	62/0	62/0	62/0	–/–	
prompt	–/–	**72**/0	59/0	59/0	59/0	–/–	
gemma-3-12b-pt							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	47/0	60/4	50/0	50/0	50/0	–/–	52/1
last	19/0	42/6	58/4	58/4	58/4	62/2	
p-wm	58/6	**67**/5	48/2	48/2	48/2	–/–	
mean	18/–	66/4	40/1	40/1	40/1	–/–	
prompt	–/–	58/**18**	44/3	44/3	44/3	–/–	
gemma-3-12b-it							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	40/**3**	**75**/0	71/0	71/0	71/0	–/–	57/0
last	57/1	55/-1	64/0	64/0	64/0	9/-1	
p-wm	68/0	72/-2	52/-2	52/-2	52/-2	–/–	
mean	24/–	71/-2	49/-1	49/-1	49/-1	–/–	
prompt	–/–	71/-2	51/-4	51/-4	51/-4	–/–	
Ministral-3-14B-Base-2512							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	55/0	80/**3**	**85**/0	**85**/0	**85**/0	–/–	60/0
last	79/0	80/2	81/2	81/2	81/2	83/0	
p-wm	79/0	84/0	83/1	83/1	83/1	–/–	
mean	63/0	84/0	81/1	81/1	81/1	–/–	
prompt	–/–	82/0	77/0	77/0	77/0	–/–	
Ministral-3-14B-Instruct-2512							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–/–	**59**/0	46/**0**	46/**0**	46/**0**	–/–	45/**0**
last	44/**0**	54/**0**	49/**0**	49/**0**	49/**0**	46/**0**	
p-wm	48/**0**	57/**0**	58/**0**	58/**0**	57/**0**	–/–	
mean	39/**0**	**59**/0	57/**0**	57/**0**	57/**0**	–/–	
prompt	1/**0**	57/**0**	43/**0**	43/**0**	43/**0**	–/–	

Table 16: Validation counterpart to Table 13. The optimal convex combination weights are evaluated on the **BOUQuET** dataset. Right to the '/' symbol is the improvement against using `src-en` alone.

Qwen3-14B-Base							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	58.1	62.6	79.0	79.0	79.0	–	81.5
last	43.6	59.3	51.6	51.6	51.7	31.0	
p-wm	**91.1**	82.1	56.7	56.7	56.7	–	
mean	84.4	80.6	**91.2**	**91.2**	**91.1**	–	
prompt	–	74.5	44.2	44.2	44.1	–	
Qwen3-14B							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	80.4	83.9	83.9	83.9	–	79.0
last	45.5	62.4	54.6	54.6	54.6	72.3	
p-wm	**88.5**	81.7	56.8	56.8	56.8	–	
mean	76.9	80.8	85.0	85.0	84.9	–	
prompt	–	68.5	37.7	37.7	37.8	–	
gemma-3-12b-pt							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	32.5	50.1	52.6	52.6	52.6	–	51.4
last	19.0	56.1	64.9	64.9	63.7	45.8	
p-wm	**66.2**	46.9	52.9	52.9	52.9	–	
mean	5.2	46.6	53.0	53.0	53.0	–	
prompt	–	36.8	49.4	49.4	48.9	–	
gemma-3-12b-it							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	15.3	43.9	56.5	56.5	56.5	–	38.3
last	53.6	39.3	62.4	62.4	62.4	25.9	
p-wm	**66.8**	49.9	53.8	53.8	53.8	–	
mean	9.1	49.4	52.1	52.1	52.1	–	
prompt	–	56.0	50.4	50.4	50.4	–	
Ministral-3-14B-Base-2512							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	56.5	80.8	86.4	86.4	86.4	–	67.2
last	82.0	84.0	85.6	85.6	85.6	80.7	
p-wm	78.3	84.2	**88.8**	**88.8**	**88.9**	–	
mean	60.9	83.9	87.6	87.6	87.7	–	
prompt	–	78.0	70.9	70.9	70.9	–	
Ministral-3-14B-Instruct-2512							
	cos	ANC	abs	dist	ratio	Dali	Eflo
fewshot	–	75.4	65.0	65.0	65.0	–	65.6
last	58.5	71.6	64.4	64.4	64.4	70.2	
p-wm	66.5	75.1	**81.1**	**81.1**	**81.1**	–	
mean	49.3	76.1	79.3	79.3	79.2	–	
prompt	6.2	73.5	58.4	58.4	58.4	–	

Table 17: Validation counterpart to Table 14. The optimal convex combination weights are evaluated on the **BOUQuET** dataset.

Figure 7: Comparison of how translation PMI (left) and CHRF (right) correlate with the CLA scores in the best-case scenario for the **Qwen3** family. Each point is a language pair; non-Latin-script target languages are in orange. Both metrics are correlated with their respective optimal convex combination (see Section 6.3) of `src-en`, `en-tgt`, and `src-tgt` CLA scores, using the best-correlating sentence representation and CLA metric. The optimal convex combination weights, representations, and metrics were calculated on the **Flores** dataset (odd rows) and reused on the **BOUQuET** dataset (even rows) for validation.

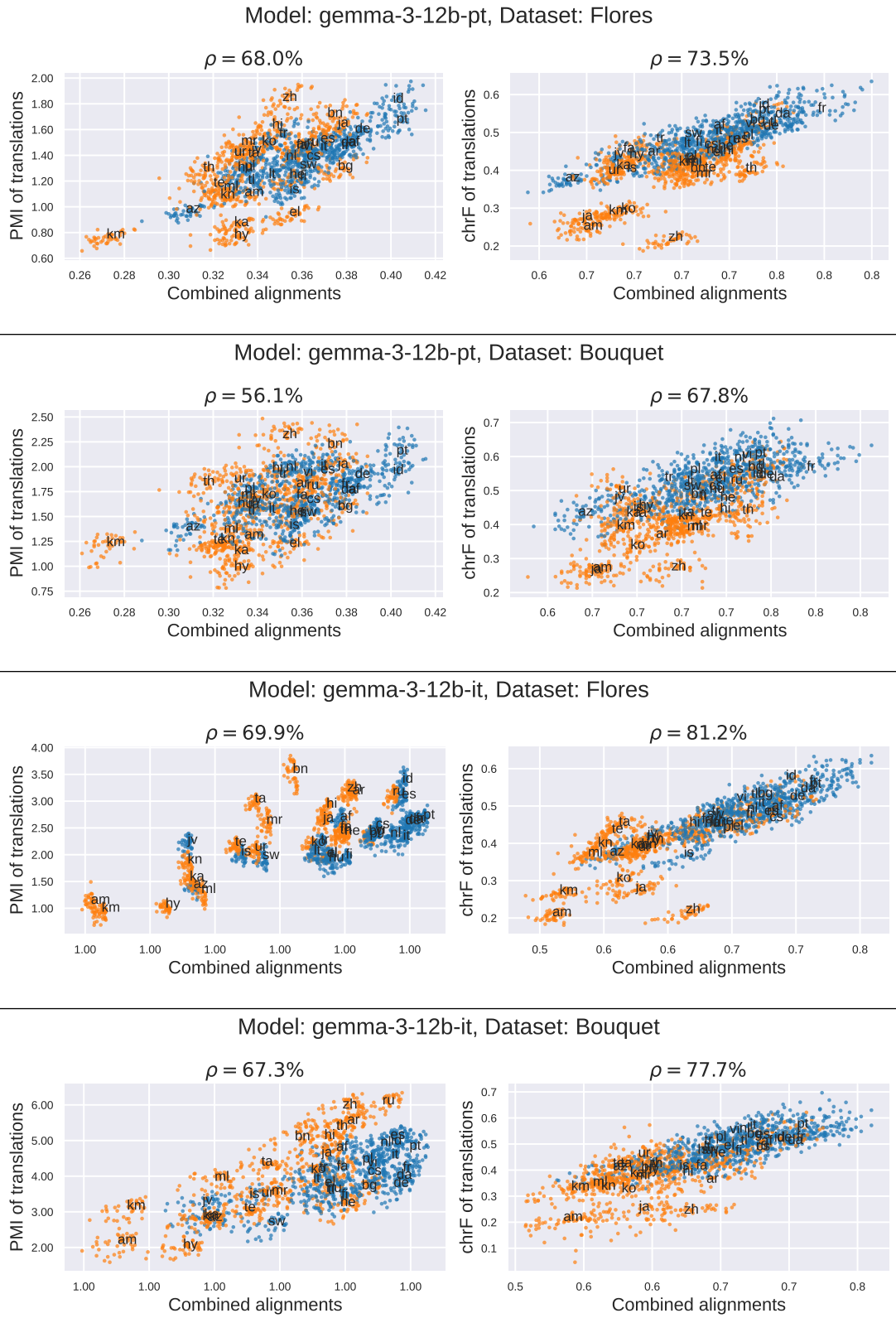

Figure 8: Comparison of how translation PMI (left) and CHRF (right) correlate with the CLA scores in the best-case scenario for the **gemma-3** family. Each point is a language pair; non-Latin-script target languages are in orange. Both metrics are correlated with their respective optimal convex combination (see Section 6.3) of `src-en`, `en-tgt`, and `src-tgt` CLA scores, using the best-correlating sentence representation and CLA metric. The optimal convex combination weights, representations, and metrics were calculated on the **Flores** dataset (odd rows) and reused on the **BOUQuET** dataset (even rows) for validation. **Note:** the outlying `gemma-3-12b-it` PMI performance on Flores is an anomaly observed in Table 4, where it correlated weakly with CHRF, compared to other models. This further leads to wrong parameter estimation and poor validation performance on BOUQuET.

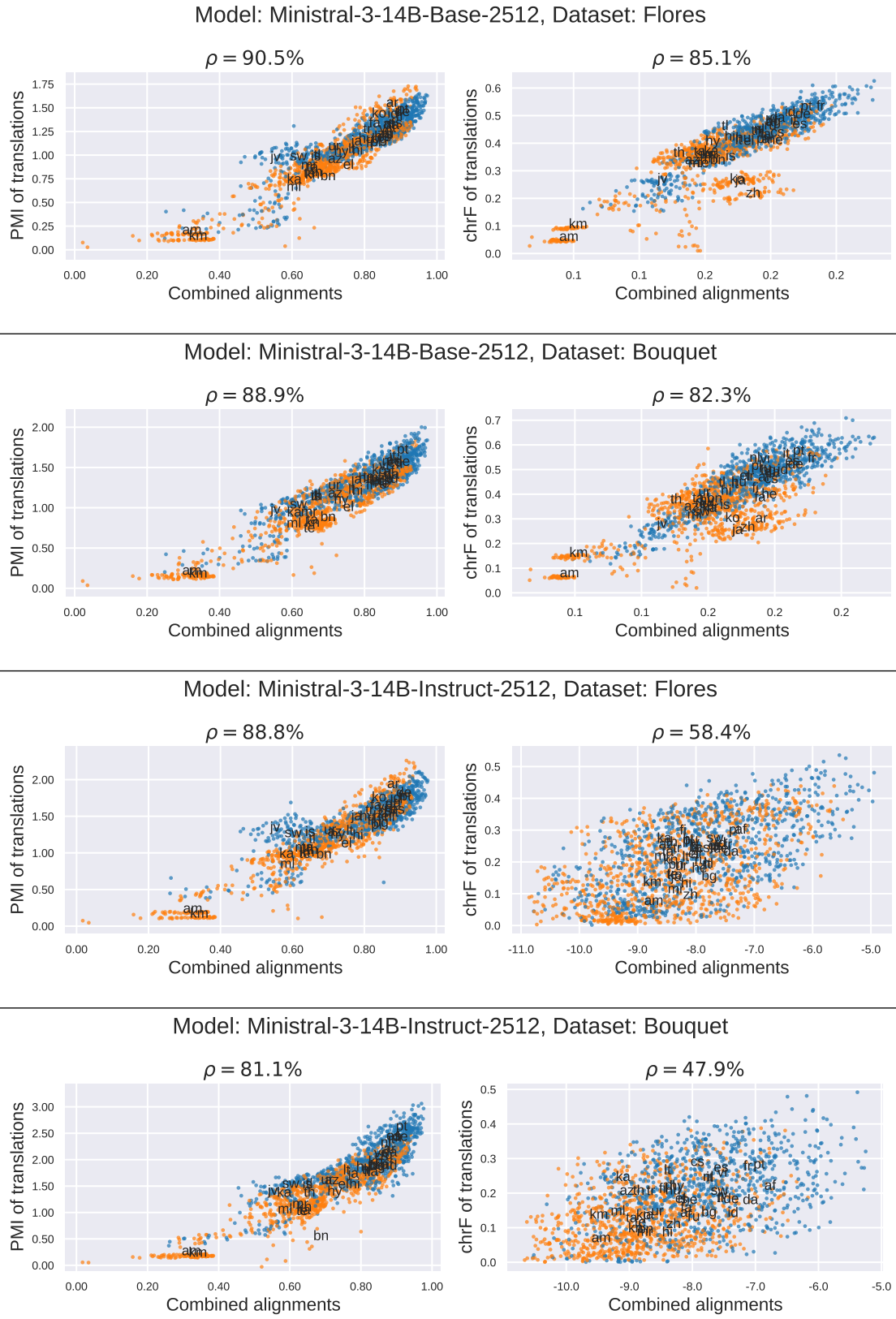

Figure 9: Comparison of how translation PMI (left) and CHRF (right) correlate with the CLA scores in the best-case scenario for the **`Ministral-3`** family. Each point is a language pair; non-Latin-script target languages are in orange. Both metrics are correlated with their respective optimal convex combination (see Section 6.3) of `src-en`, `en-tgt`, and `src-tgt` CLA scores, using the best-correlating sentence representation and CLA metric. The optimal convex combination weights, representations, and metrics were calculated on the **Flores** dataset (odd rows) and reused on the **BOUQuET** dataset (even rows) for validation. **Note** that the noise in CHRF for the instr. variant is likely caused by the model often failing to follow the translation instructions, as described in Section 6.2.