Can you map it to English? The Role of Cross-Lingual Alignment in the Multilingual Performance of LLMs

Kartik Ravisankar **HyoJung Han** **Sarah Wiegreffe** **Marine Carpuat**
University of Maryland, College Park, MD, USA
{kravisan, hjhan, sarahwie, marine}@umd.edu

Abstract

Large language models (LLMs) can answer prompts in many languages, despite being trained predominantly on English; yet, the mechanisms driving this generalization remain poorly understood. This work asks: How does an LLM's ability to align representations of non-English inputs to English impact its performance on natural language understanding (NLU) tasks? We study the role of representation alignment in instance-level task decisions, complementing prior analyses conducted both at the language level and task-independently. We introduce the Discriminative Alignment Index (DALI) to quantify instance-level alignment across 24 languages other than English and three distinct NLU tasks. Results show that incorrect NLU predictions are strongly associated with lower representation alignment with English in the model's middle layers. Through activation patching, we show that incorrect predictions in languages other than English can be fixed by patching their parallel English activations in the middle layers, thereby demonstrating the causal role of representation (mis)alignment in cross-lingual correctness.[1]

1 Introduction

Large language models (LLMs) exhibit impressive multilingual capabilities, successfully performing natural language understanding (NLU) tasks such as reading comprehension and common-sense reasoning in languages other than English despite being pre-trained mostly on English text (Touvron et al., 2023; Muennighoff et al., 2023). This ability to transfer NLU capabilities from high-resource languages (e.g., English) to lower-resource ones has been well-documented in encoder-only architectures (Conneau et al., 2018, 2020; Yang et al., 2019; Devlin et al., 2019). However, the capacity of decoder-only LLMs to internalize and transfer knowledge across languages remains relatively underexplored (Hämmerl et al., 2024).

Recent work has focused on understanding how LLMs predominantly trained in English process multilingual text. Wendler et al. (2024) analyze intermediate representations in Llama-2 (Touvron et al., 2023) through early exit strategies (nosalgebraist, 2020), demonstrating an implicit pivot through English in the middle layers while processing non-English text. This raises the question of whether an LLM's ability to align representations of non-English text to its corresponding parallel English text representations is predictive of its capabilities in languages other than English.

Prior work shows that representational alignment measured using independent parallel corpora correlates with multilingual task accuracy: Kargaran et al. (2025) introduced MEXA, a cross-lingual retrieval-based score for languages other than English, computed by comparing non-English representations to English representations from the same model across a fixed set of parallel texts. The authors demonstrate that MEXA scores are strongly correlated with multilingual task accuracy (i.e., languages with better MEXA scores exhibit better task accuracy), suggesting that cross-lingual alignment is a good indicator of an LLM's multilingual capability. However, language-level correlational studies analyze alignment at the aggregate level, using a corpus independent of the NLU task, leaving open the question of how representation alignment affects model behavior on individual NLU instances.

Our work seeks to address this gap and determine whether LLMs make better multilingual predictions when given inputs well-aligned with English. We ask two research questions (RQ):

1. **RQ1:** Is cross-lingual alignment associated with task accuracy at the instance-level of a specific NLU task within a language?

2. **RQ2:** Does misalignment with English repre-

[1] **Code:** github.com/Kartik21/XLingAlignment

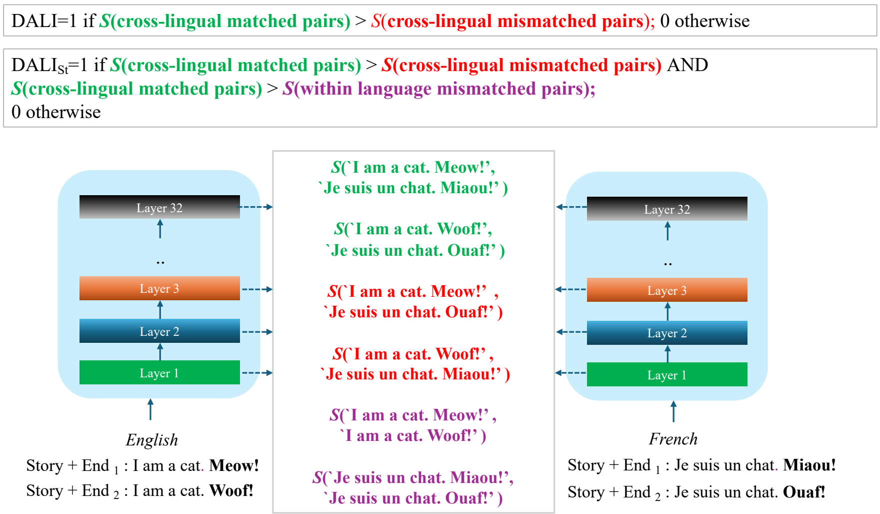

DALI=1 if S(cross-lingual matched pairs) > S(cross-lingual mismatched pairs); 0 otherwise
DALIst=1 if S(cross-lingual matched pairs) > S(cross-lingual mismatched pairs) AND
S(cross-lingual matched pairs) > S(within language mismatched pairs);
0 otherwise
Layer 32
..
Layer 3
Layer 2
Layer 1
English
Story + End 1 : I am a cat. Meow!
Story + End 2 : I am a cat. Woof!
S(`I am a cat. Meow!', `Je suis un chat. Miaou!')
S(`I am a cat. Woof!', `Je suis un chat. Ouaf!')
S(`I am a cat. Meow!', `Je suis un chat. Ouaf!')
S(`I am a cat. Woof!', `Je suis un chat. Miaou!')
S(`I am a cat. Meow!', `I am a cat. Woof!')
S(`Je suis un chat. Miaou!', `Je suis un chat. Ouaf!')
Layer 32
..
Layer 3
Layer 2
Layer 1
French
Story + End 1 : Je suis un chat. Miaou!
Story + End 2 : Je suis un chat. Ouaf!

Figure 1: DALI is calculated at the instance-level across transformer layers using its representations. The model is tasked with picking the right ending (*'Meow/Woof'* in English; *'Miaou/Ouaf'* in French) given a premise (*'I am a cat/Je suis un chat.'* in English and French, respectively). DALI=1 if the cosine similarity S of both the cross-lingual matched pairs exceeds both the mismatched pairs, indicating the ability of the model to distinguish semantically equivalent English and French text from non-equivalent representation pairs. A stricter variant, DALI$_{st}$ adds an additional condition that S of cross-lingual matched pairs must exceed the intra-lingual mismatched pairs.

sentations causally explain failures on inputs in other languages?

Following Kargaran et al. (2025), we focus on multilingual discriminative NLU tasks. To answer RQ1, we introduce the Discriminative Alignment Index score (DALI) and its variant (DALI$_{st}$), which quantify representation alignment at the instance level (Figure 1). By establishing instance-level metrics to quantify alignment, we compare two groups in which cross-lingual transfer from English yields distinct outcomes: **Transfer Success (TS)** and **Transfer Failure (TF)** (Figure 2). We find that TS instances are consistently more aligned with their English counterparts than TF instances across benchmarks, languages, and models.

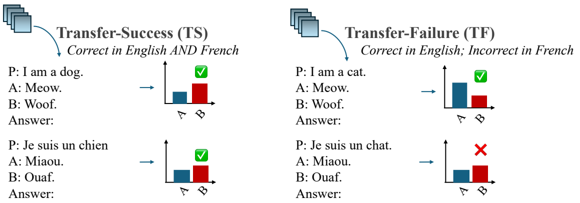


Figure 2: French TS and TF samples in MCQA format.

We answer RQ2 through controlled activation patching (Figure 3): we patch semantically equivalent English representations at layer λ onto the corresponding non-English forward pass of TF instances. We then observe whether the patch is more successful at flipping the model's prediction to the correct answer than a patch from a control (i.e., semantically non-equivalent) instance. Our experiments reveal that semantically equivalent English patches are consistently more effective than control patches at correcting predictions, with effects concentrated in specific intermediate layers, providing causal evidence that alignment with English representations drives successful cross-lingual transfer.

2 Methods

We first introduce metrics to quantify representation alignment with English at an instance level, and compare alignment between TS and TF instances (§2.2). Then, we formalize our activation patching setup, which establishes a causal link between alignment and cross-lingual transfer (§2.3).

2.1 Preliminaries

MEXA (Kargaran et al., 2025), henceforth denoted as MEXA$_F$, measures a model's general cross-lingual alignment ability with English using a fixed set of sentences from parallel datasets such as FLORES-200 (Team et al., 2022). MEXA$_F$ follows the concept of weak alignment (Hämmerl et al., 2024) defined as the proportion of samples that are *'aligned'*. For languages $\mathrm{L}_1, \mathrm{L}_2$ drawn from the set $\mathcal{L}$, let (u_i, v_i) be pairs of sentence representations derived from the layer λ of a transformer where $i = 1, \ldots, N; u \in \mathrm{L}_1, v \in \mathrm{L}_2$. We say a sample is *'aligned'* if it has a higher cosine-similarity with its parallel instance than with other non-parallel instances. Formally, we define $\texttt{MEXA}_F$ as follows:

$$\texttt{MEXA}_{F(\mathrm{L}_1,\mathrm{L}_2,\lambda)} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big(S(u_i,v_i) > \max_{j\in 1,\dots N; j\neq i}\big(\{S(u_i,v_j)\}\cup\{S(u_j,v_i)\}\big)\big) \tag{1}$$

Since transformer representations exhibit anisotropy (Ethayarajh, 2019), i.e., they occupy a narrow, directional cone in the latent space rather than being uniformly distributed, cosine similarity scores are high even for semantically unrelated text, making it challenging to distinguish genuine alignment from spurious directional clustering. By assigning a binary score instead of using raw cosine similarities, $\texttt{MEXA}_F$ mitigates this issue. Any parallel dataset can be used to compute $\texttt{MEXA}_F$, as evidenced by the original study, which also used the Bible (Mayer and Cysouw, 2014) corpus in addition to FLORES-200. Empirically, $\max_\lambda\{\texttt{MEXA}_{F(\mathrm{L}_1=\mathrm{English},\mathrm{L}_2)}\}$ is strongly correlated with NLU model performance in L_2, such that languages with higher $\texttt{MEXA}_F$ scores tend to achieve higher task accuracy and vice versa.

2.2 Instance-level metrics for Cross-lingual Alignment

While $\texttt{MEXA}_F$ provides a general magnitude of cross-lingual alignment at the language-level, our objective is to quantify cross-lingual alignment at an instance-level on a discriminative task. To this end, we propose three metrics: 1) $\texttt{DALI}$, 2) $\texttt{DALI}_{st}$ - a stricter variant of $\texttt{DALI}$, and 3) $\texttt{MEXA}_T$ - a task-specific variant of $\texttt{MEXA}_F$. Although these metrics can be applied to any two languages, we fix L_1 = English for all our analyses, as we are interested in examining the LLM's ability to align non-English representations with their corresponding English counterparts.

DALI Consider a discriminative task $\mathcal{D}$ across multiple languages, where each instance has a premise P and string answer options $o_1, o_2, \dots, o_{n_{\text{opt}}}$. The model is tasked with picking the correct option. Figure 1 presents one example where the model is given a premise in both English (P_{Eng} = *'I am a cat.'*) and French (P_{Fr} = *'Je suis un chat.'*), and $n_{\text{opt}} = 2$. The model is tasked with picking the right ending among the options for the given premise, conditioned on the language of the premise: in English, $o_{1,\text{Eng}}$ = *Meow!* and $o_{2,\text{Eng}}$ = *Woof!*; in French, $o_{1,\text{Fr}}$ = *Miaou!* and $o_{2,\text{Fr}}$ = *Ouaf!*.

We extract the representations of the premise-ending combinations $(P + o_1, P + o_2)$ in both languages from various layers λ of an LM. We set $\texttt{DALI}$ = 1 if the similarity (S) of cross-lingual matched pairs exceeds mismatched pairs. In principle, we can use any vector similarity metric, but we use cosine similarity as a representative option. $\texttt{DALI}$ thus captures the model's ability to align parallel premise and ending representations of English and non-English samples. We define $\texttt{DALI}$ for a given sample across languages $\mathrm{L}_1, \mathrm{L}_2$ based on the representations in layer λ of a transformer as follows:

$$\texttt{DALI}_{\mathrm{L}_1,\mathrm{L}_2,\lambda} = \begin{cases} 1, \text{if } S\big((P+\mathrm{o}_i)_{\mathrm{L}_1},(P+\mathrm{o}_i)_{\mathrm{L}_2}\big) \\ \quad > S\big((P+\mathrm{o}_i)_{\mathrm{L}_1},(P+\mathrm{o}_j)_{\mathrm{L}_2}\big) \\ \quad \forall i,j = 1,\dots,n_{\text{opt}};\ i\neq j \\ 0, \text{otherwise} \end{cases} \tag{2}$$

We follow the same approach as $\texttt{MEXA}$ (Kargaran et al., 2025) by assigning a binary $\texttt{DALI}$ score per sample instead of cosine similarities. However, the small pool of mismatched pairs reduces $\texttt{DALI}$'s discriminative power: for instance, a 2-option task involves only two cross-lingual mismatches, increasing the likelihood of false positives.

$\texttt{DALI}_{st}$ To address this issue, we introduce a stricter metric, $\texttt{DALI}_{st}$. We enforce an additional criterion that the cosine similarity of cross-lingual matched pairs must surpass all within-language mismatched pairs (see Figure 1). This imposes a stricter threshold by capturing whether the similarity between cross-lingual pairs is higher than that of non-equivalent sentences within the same language. We formally define $\texttt{DALI}_{st}$ as follows:

$$\texttt{DALI}_{st,\mathrm{L}_1,\mathrm{L}_2,\lambda} = \begin{cases} 1, \text{if } \texttt{DALI} = 1 \text{ and} \\ S\big((P+\mathrm{o}_i)_{\mathrm{L}_1},(P+\mathrm{o}_i)_{\mathrm{L}_2}\big) \\ \quad > S\big((P+\mathrm{o}_i)_{\mathrm{L}_k},(P+\mathrm{o}_j)_{\mathrm{L}_k}\big) \\ \quad \forall i,j = 1,\dots,n_{\text{opt}}; i\neq j, \forall k = 1,2 \\ 0, \text{otherwise} \end{cases} \tag{3}$$

$\texttt{MEXA}_T$ We also compute a task-specific version of $\texttt{MEXA}_F$ (Equation 1) that allows for instance-level alignment measurements for each NLU sample. The only difference in calculating $\texttt{MEXA}_T$ is that $u \in P_{\text{Eng}}, v \in P_{\mathrm{L}_2}$ from task $\mathcal{D}$ as opposed to being generic sentences from the FLORES dataset.

Benchmark	Description	$\lvert\mathcal{L}\rvert$	n_{opt}	N
Belebele (Bandarkar et al., 2024)	Reading comprehension	122	4	900
Xstorycloze (Lin et al., 2022)	Narrative understanding	11	2	1511
Xcopa (Ponti et al., 2020)	Commonsense reasoning	11	2	500

Table 1: NLU discriminative benchmarks– Number of languages ($|\mathcal{L}|$), options per sample (n_{opt}), and number of samples per language (N).

Differences between the metrics While the three proposed metrics all quantify cross-lingual alignment at the instance level, they differ in their definitions. DALI and DALI$_{st}$ ensure that premise+option pairs in a given NLU instance are aligned, meaning the contrastive examples are within a given test instance. MEXA$_T$, on the other hand, focuses only on whether the representations of parallel premises across languages are more aligned than non-parallel premises. Based on Figure 1's example,

$$\texttt{MEXA}_T = \begin{cases} 1 & \text{if } S(\text{`Je suis un chat'}, \text{`I am a cat'}) \\ & \quad > \max\big\{S(\text{`I am a cat'}, \text{other } P_{\text{Fr}}), \\ & \qquad S(\text{`Je suis un chat'}, \text{other } P_{\text{Eng}})\big\} \end{cases} \quad (4)$$

Due to the number of contrastive parallel samples ($2N-2$ if there are N total samples), the probability of MEXA$_T = 1$ occurring by chance is low. On the other hand, DALI relies on within-sample mismatched pairs, which are inherently limited by task design: a 2-option task involves only two mismatched cross-lingual pairs. While DALI$_{st}$ mitigates this by enforcing a stricter criterion, tasks with few options remain vulnerable to false positives due to anisotropy (§5.1).

Associative analysis of cross-lingual alignment We analyze the role of alignment between **TS** and **TF** instances for each language. We hypothesize that TS samples exhibit *higher* cross-lingual alignment than TF samples, and we test this hypothesis statistically. It is worth noting that DALI, DALI$_{st}$, and MEXA$_T$ are derived across all transformer layers for each sample, producing a binary vector (of the same length as the number of layers) per instance. We compute the % of samples with alignment=1 at each layer λ and identify the layer with the largest alignment overall (denoted as λ_{max}). Although statistical tests between alignment measures can be conducted at any layer, we localize our tests to the layer with the highest alignment for each metric, as this is where the model's cross-lingual alignment mechanism is most actively and successfully engaged. We compare alignment (% alignment=1 at λ_{max}) across the two groups using a z-test for proportions against a one-sided alternative that % aligned in TS samples exceeds % aligned in TF samples at a level (α) of 0.05. A positive $\Delta_{\text{TS-TF}}$(alignment) means that samples that generalize well have a higher degree of alignment than samples that do not.

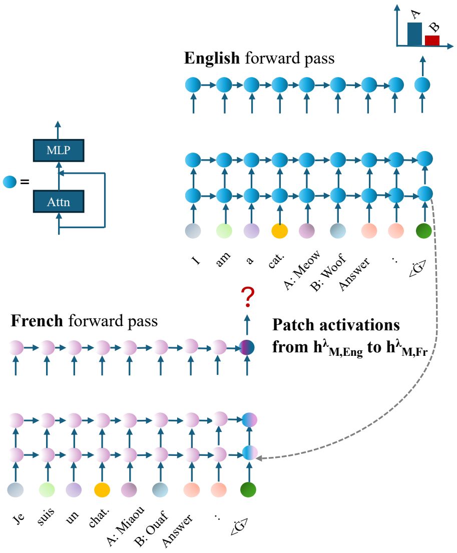


Figure 3: We analyze the causal effect of alignment on task accuracy by patching a successful English forward pass $h^{\lambda}_{M,Eng}$ to the unsuccessful French forward pass at layer λ ($h^{\lambda}_{M,Fr}$), thereby treating h^{λ}_{Eng} as a causal mediator for the French sample's success. We measure the effectiveness of the patch by tracking the % of samples that flip to the correct prediction upon patching.

2.3 Activation Patching

To determine whether non-English failures stem from the model's inability to construct representations equivalent to those produced by the English forward pass, we perform activation patching. Activation patching is an interpretability technique (Vig et al., 2020; Meng et al., 2022) that allows one to test whether an LM's representation at layer λ (h^{λ}) is a causal mediator (Mueller et al., 2024) for the model's behavior. Transformer architectures process input tokens[2] $x_1, x_2, \ldots, x_m, \ldots, x_M \in \mathcal{V}$ and transform each token x_m into the d-dimensional embedding space at layer 0 (h^0_m). For each token and layers $\lambda \in$

[2] Without loss of generality, we assume the input tokens are specific to language L.

$\{1, \ldots, \Lambda\}$, the transformer model updates the residual stream in the following way, where $f_\lambda(.)$ refers to the self-attention block followed by the multi-layer perceptron (MLP) block at layer λ:

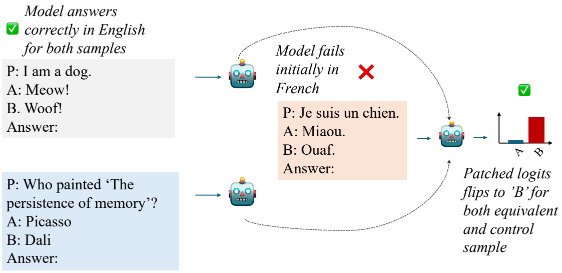
$$h^{\lambda}_{m,L} = h^{(\lambda-1)}_{m,L} + f_\lambda(h^{(\lambda-1)}_{1,L}, \ldots, h^{(\lambda-1)}_{m,L}) \quad (5)$$

The next-token logits are derived by unembedding the vector output by the final layer, $h^{\Lambda}_{m,L}$, back to the vocabulary space $\mathcal{V}$. For simplicity, we drop the token index subscript m in subsequent sections, noting that patching is performed only at specific positions such as the last or penultimate token.

We run two forward passes for each TF sample: one in English, which the model answers correctly, and one in another language, which the model answers incorrectly. If the non-English forward pass had constructed a representation at layer λ equivalent to h^{λ}_{Eng}, would the non-English prediction have instead been correct? We test this by copying the activations at specific token positions from the English forward pass (h^{λ}_{Eng}), and patching them onto the corresponding non-English forward pass at the same layer (h^{λ}_{X}) (see Figure 3). We iterate this process across all layers of the transformer, enabling us to localize the layers where aligned representations are sufficient to correct failures in other languages.

Unlike in §2.2's experiments, which computed alignment by concatenating the premise and options, we frame the discriminative NLU tasks as Multiple-Choice Question Answering (MCQA) to ensure that key token positions align across the two instances (a prerequisite for patching). Formally, we have a set of target tokens $\mathcal{Y} = \{y_1, y_2, \ldots, y_{n_{\text{opt}}}\}$ (e.g., $\{\texttt{A}, \texttt{B}\}$ in the 2-choice MCQA task) that represent string answer choices $o_1, o_2, \ldots, o_{n_{\text{opt}}}$ for each instance. We use the same Latin-script target tokens across languages. The model's prediction is considered correct if the logit for the target token representing the correct answer (y_c) is higher than those for the other target tokens (Refer Figure 2 for an example).

Control In an MCQA setting, models may encode either the predicted answer choice string or its corresponding answer symbol token (or both) in their hidden states, depending on the layer (Wiegreffe et al., 2025). If h^{λ}_{Eng} encoded the target token prediction rather than *only* information specific to the instance, an unrelated English sample with the same y_c (e.g., '`A`') would be equally effective at correcting the non-English prediction. To isolate the layers and token positions where h^{λ}_{Eng} encodes concepts or semantics and not simply the target token or pointers to it, we also perform control patching where we copy activations associated with unrelated English samples, but with the same y_c as the current instance pair. In Figure 4, we demonstrate patching h^{λ}_{Eng} from an unrelated sample with $y_c =$ '`B`' to repair the French sample. The successful control patch demonstrates that h^{λ}_{Eng} cannot be interpreted as encoding *only* semantic information and is thus inconclusive for testing cross-lingual alignment for this particular instance. The control experiment exemplifies why we frame the tasks as MCQA: it provides the flexibility to evaluate and identify key, consistent token positions despite differences in tokenization across languages.


✓ Model answers correctly in English for both samples
Model fails initially in French ✗
✓
A
B
Patched logits flips to 'B' for both equivalent and control sample

P: I am a dog.
A: Meow!
B. Woof!
Answer:

P: Je suis un chien.
A: Miaou.
B: Ouaf.
Answer:

P: Who painted 'The persistence of memory'?
A: Picasso
B: Dali
Answer:

Figure 4: Control experiment to check if h^{λ}_{Eng} encodes semantic concepts or just the target token.

Token position A key facet of our experiment is the token position m from which we must copy our English activations and patch them into the non-English forward pass. We consider two positions in our Multiple-Choice Question Answering (MCQA) prompts—at the **penultimate token** $(M-1)$ and the **last token** (M), since both positions capture the task logic of the entire sequence across languages.

Metrics We evaluate the causal effectiveness of English representations on non-English instances in three dimensions:

1. **Mean logits** over the non-English TF samples, identifying the layer λ where the logit associated with the gold token (y_c) exceeds the logits assigned to the distractor target tokens,

2. **% flips**, defined as the proportion of non-English TF samples where the y_c's logit exceeds the distractor logits post-patching, including the logit of the incorrect token initially predicted pre-patching, and

3. **Δ % flips** between equivalent and control patching setups, since this isolates the role of conceptual alignment from simple target-token production.

3 Experimental Setup

Tasks We evaluate LLMs on three multilingual NLU benchmarks (Table 1). Our study is made possible by their parallel nature, in which premise-option pairs are structurally identical and equivalent across languages (e.g., 'I am a cat' in English and 'Je suis un chat' in French).

The specific prompts used are detailed in Figure 10. For each sample, we use the MCQA format and evaluate the logits for each target token $y \in \mathcal{Y}$. We consider the model's decision correct if the gold token (y_c) receives the highest logits. We label an instance as TS if the model is correct on both the English and the non-English NLU sample, and TF if it is correct only on the English version (Figure 2).

Models and Languages We conduct our alignment association experiments across three diverse open-source models, namely Llama3.1 8B , Llama3.1 8B (it) (Grattafiori et al., 2024), and Aya23 8B (Aryabumi et al., 2024). This selection encompasses diverse architectural approaches, supported languages, and training methodologies. Critically, these models report and document the range of languages they support, which enabled us to construct an intersection set with the languages available in individual benchmarks (Table 3).

Since the patching experiments are done layer-wise, this yields 2Λ forward passes per TF sample in a given language. Due to computational constraints, we limit our patching experiments to the Llama3.1 8B model and the languages it natively supports. We perform our patching experiments using `nnsight` (Fiotto-Kaufman et al., 2025), a Python package used to access model internals.

Embeddings Following prior work (Neelakantan et al., 2022; Wang et al., 2024; Kargaran et al., 2025; Li et al., 2025), we extract the embeddings corresponding to the last token of the text (h^{λ}_{M}) across each layer of the transformer to calculate our instance-level alignment metrics.

4 Results

We present the results of our instance-level alignment analysis (§4.1), the activation patching (§4.2), and the control patching experiments (§4.3).

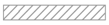
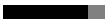
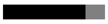
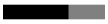
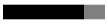
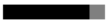
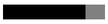
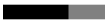
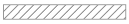
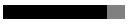
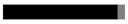
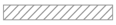
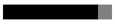
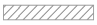
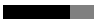
Benchmark	Language	n_{acc}	$\%_{acc}$	TS/TF distribution
Belebele	English	723	80.3	
	French	648	72.0	
	German	630	70.0	
	Hindi	499	55.4	
	Italian	624	69.3	
	Portuguese	655	72.8	
	Spanish	635	70.6	
	Thai	499	55.4	
Xstorycloze	English	1451	96.0	
	Hindi	1263	83.6	
	Spanish	1398	92.5	
Xcopa	English-Italian	429	85.8	
	Italian	408	81.6	
	English-Thai	356	71.2	
	Thai	288	57.6	

Table 2: Accuracy results for Llama3.1 8B. n_{acc} and $\%_{acc}$ present the number of accurate instances in a given language. The distribution column presents the % of accurate samples in English (▨) for each benchmark, as well as the breakdown of TS (■) and TF (▒) instances for each language-benchmark combination.

4.1 Is alignment stronger when transfer succeeds?

Table 2 presents the accuracies achieved by the Llama3.1 8B model for each language-benchmark combination, along with the % of TS and TF samples used in our alignment experiments.

The $\Delta_{\text{TS-TF}}$(alignment) is consistently positive (Figure 5) in 30 out of the 33 (language × benchmark × alignment metric) combinations, including 8 with statistically significant differences. This finding directly addresses RQ1 as we demonstrate an association between cross-lingual alignment with English and instance-level success. Considering the various aspects of cross-lingual alignment that these metrics measure (§2.2), the consistently positive Δs point towards a strong association between superior representational alignment and successful transfer generalization. We observe similar positive results in other models (Figure 11).

4.2 Does patching $h^{\lambda}_{\textbf{Eng}}$ correct failures?

We present the results of the Llama3.1 8B semantically equivalent patching experiment on the Belebele benchmark for Italian (148 TF samples) in the last token (Figure 6a) and the penultimate token (Figure 6b) settings. Results for other language-benchmark combinations are presented in Figures 12, 13, and 14. We divide the plots into three regions of the model's decision process.

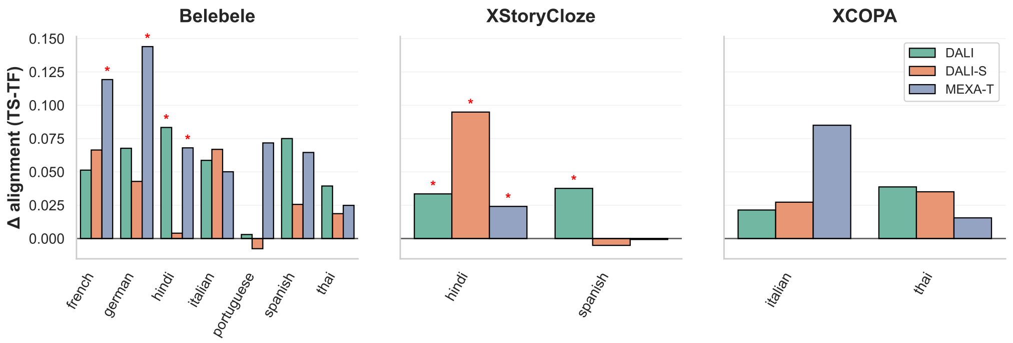
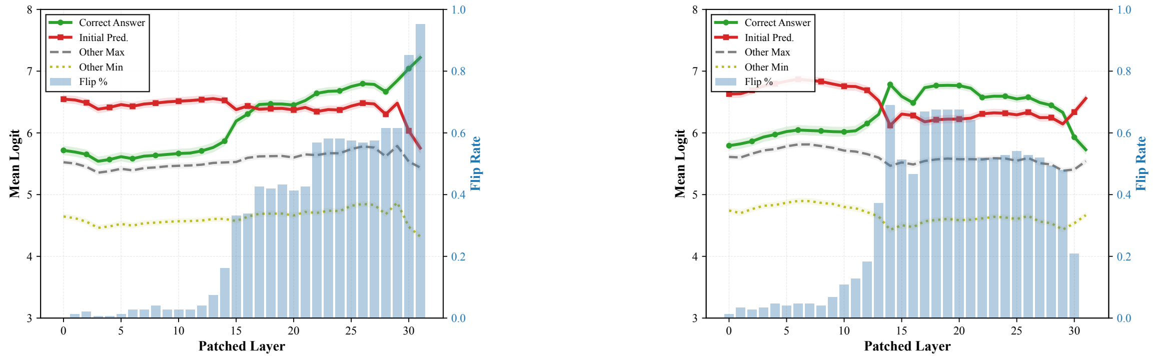


Figure 5: Δ Alignment between TS and TF samples using DALI, DALI$_{st}$, and MEXA$_T$ across the three benchmarks in Llama 3.1 8B. Positive Δs indicate that alignment is higher in TS samples than TF samples for a given language. The asterisk (*) indicates that this difference is statistically significant (p < 0.05).



(a) Belebele - Italian (last-token)

(b) Belebele - Italian (penultimate token)

Figure 6: Mean logit trajectory and % flip for Italian TF samples (N=148) in the belebele benchmark: Blue bars represent the % of TF samples that flip to the correct prediction; Green line refers to the mean logits of the gold-token (t_c) and Red line refers to the mean logits of the original incorrect prediction.

Last Token - **Early layers (0-15):** % flip remains low, and the mean logits of the initial incorrect answer exceed the gold answer token, indicating that h^{λ}_{Eng} has not yet aggregated the right features necessary to overturn the wrong prediction; **Middle layers (16-25):** % flip rises sharply in layer 16 and continues to rise until layer 25. This suggests that h_{Eng} corresponding to the last token in the middle layers (as early as layer 17) encodes a language-agnostic representation that is sufficient to steer the non-English samples towards the gold token; **Later layers (25-32):** % flip steadily increases followed by maximal correction, trivially reaching ≈ 100% in the last layer (i.e., if h^{Λ}_{Eng} is patched to h^{Λ}_{Ita} at the last token, the model is essentially treating it as an English sample).

Penultimate Token - **Early layers (0-13):** % flip remains low, indicating that h^{λ}_{Eng} has not yet aggregated the right features necessary to overcome the Italian failures; **Middle layers (14-25):** This is followed by a spike % flip in layer 14 and concurrently resulting in the mean logits of gold answer token surpassing the initial wrong answer in Italian; **Later layers (25-32):** Patching h^{λ}_{Eng} in the later layers at the penultimate token is not very effective since the Italian forward pass has already executed the decision and resists changing its original incorrect prediction (indicated by the steep drop in % flip in the later layers). This behavior contrasts with patching the last token (Figure 6a), where the % flip steadily increases as we patch in the later layers, since the last-token representations are directly associated with the next-token prediction.

Overall, for both token positions, the patched h^{λ}_{Eng} in the middle layers resulted in higher mean logits associated with y_c, suggesting that transfer failures stem from the non-English forward pass constructing an *'incorrect'* representation (different from h^{λ}_{Eng}) in these layers. While patching h^{λ}_{Eng} does not flip all samples, the mean logits of the gold token surpass the initial wrong answer on average. This suggests that misalignment in the middle layers is primarily responsible for incorrect predictions for languages other than English.

4.3 What does h^{λ}_{Eng} encode?

Figure 7 shows the Δ % flip for Italian TF samples on Belebele, measuring the % of samples that flip exclusively under semantically matched patching relative to controls. We identify a select few middle layers where the Δ(% Flip) is high, indicating that the cross-lingual semantic information is causally involved in driving the model's prediction. For example, in layer 14 of the penultimate token patching (Figure 7), 41.2% (out of 68.9% total flips) can be flipped only by patching semantically equivalent English sample activation. We see this behavior consistent in other languages as well (Figure 15). While this confirms that cross-lingual conceptual alignment in the middle layers mediates correctness, it calls for future work to disentangle the effects of conceptual alignment from those of next-token production in the representations.

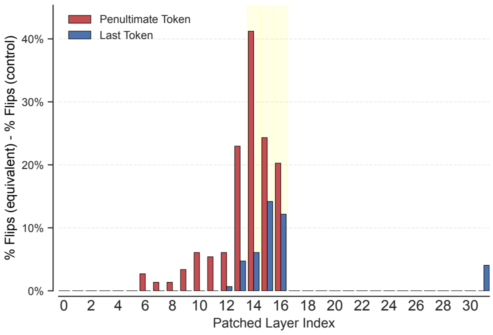


Figure 7: Control patching results - belebele Italian TF samples.

5 Discussion

Beyond our core analysis, we explore several secondary results that clarify the nature of cross-lingual representations. This includes an instance-level qualitative analysis (§5.1), an investigation into language-specific trends (§5.2), and an assessment of model entropy (§5.3).

5.1 Qualitative analysis

We illustrate a false positive instance misclassified by DALI from the XStorycloze Hindi benchmark. At layer 14, 92.9% of the samples have DALI = 1, but only 64.9% have DALI_{st} = 1, indicating false positives of alignment in DALI, but corrected by the intra-lingual constraint in DALI_{st}.

One such example ($\text{DALI} = 1; \text{DALI}_{st} = 0$) is shown in Figure 8. Patching experiments reveal that the instance flips only when patching the last-token activation at layer 30. No correction is observed at earlier layers, indicating that the Hindi forward pass remains corrupted and is only corrected by a next-token decision signal, rather than by semantic alignment. This sample was classified as 'aligned' by DALI, but *caught* by DALI_{st}, illustrating how the proposed alignment metrics differ.

Premise: कैरोलाइन, मेडिकल स्कूल की छात्रा थी. कैरोलाइन अच्छे ग्रेड पाने के लिए बहुत कड़ी मेहनत करती थी. एक दिन कैरोलाइन किसी टेस्ट में एक प्वॉइंट से फ़ेल हो गई. कैरोलाइन को बहुत हताशा हुई, लेकिन उसने कड़ी मेहनत के साथ पढ़ना जारी रखा / Caroline was a student in medical school. Caroline worked very hard to get good grades. One day Caroline failed a test by one point. Caroline was very frustrated but she continued to study hard.

Ending 1: लेकिन उसने हार मान ली. / But she gave up.
Ending 2: बाद में, वह टेस्ट में पास हो गई. / Later, she passed the test.

Figure 8: TF Hindi instance in XStorycloze.

5.2 Language-specific trends of alignment

We analyzed the % of samples with alignment = 1 ($\text{DALI}, \text{DALI}_{st}, \text{MEXA}_T$) across the LLM layers (Figure 16). We found a consistent pattern: languages that are typologically closer to English and HR languages (e.g., Spanish, German) exhibit higher alignment than languages that are not (e.g., Hindi, Thai). We also found that cross-lingual alignment peaks in the middle layers, consistent with prior works (Kargaran et al., 2025; Wilie et al., 2025; Liu and Niehues, 2025).

5.3 Entropy of equivalent and control patches

Both semantically matched, and control patches induce prediction flips in the TF examples (Figure 7 - only some layers have non-zero Δs). A *flip* to a correct answer in a discriminative instance is binary, and we investigate this further from the perspective of the entropy of the model's decision.

$$\text{entropy} = -\sum_{t \in T} p_t \times log(p_t) \qquad (6)$$

We observe (Figure 9) that the flips triggered by semantically equivalent English patches are consistently associated with lower output entropy. This suggests that semantic alignment not only increases the likelihood of correction but also stabilizes the model's decision, yielding more confident predictions than control patches that primarily inject a next-token bias. This analysis highlights an underexplored area of the effect of cross-lingual alignment on confidence calibration.

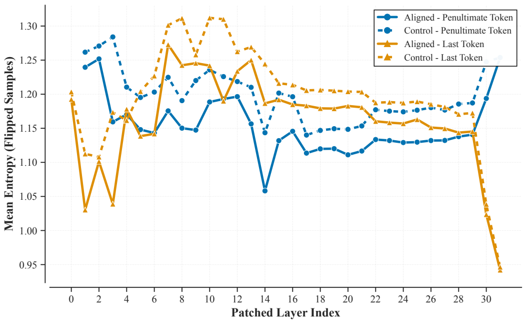


Figure 9: Entropy analysis of belebele French samples.

6 Related Work

Multilingual LLMs Multilingual LLMs are designed to process and generate text across multiple languages. However, the pretraining corpus of most state-of-the-art LLMs is dominated by English, despite exhibiting reasonable capabilities in other languages (Ahia et al., 2023). Etxaniz et al. (2023) provided evidence that multilingual LLMs perform better in English through *'self-translate'*, where LLMs were first instructed to translate the other-language prompts to English and process them in English. Wendler et al. (2024) extended this by decoding middle layer residuals to show that LLMs place more probability mass on English tokens than on tokens of the prompted language before transitioning to the target-language vocabulary in the final layers – a pattern which is interpreted as reasoning in an abstract, language-agnostic concept space in the middle layers biased toward English rather than reliance on English as an explicit lexical pivot. The presence of a language-agnostic concept space was validated by Dumas et al. (2025) using activation patching. Multiple concurrent studies provided evidence for an implicit translation $\rightarrow$ task-solving $\rightarrow$ translation pipeline underlying improved performance in languages other than English, positing through early decoding that models *'solve'* the task by decoding the right English token in the middle layers but often fail to faithfully translate the resulting *'gold'* token back into the target language during final decoding (Bafna et al., 2025; Lu et al., 2025; Wang et al., 2025a). Li et al. (2025) and Kargaran et al. (2025) demonstrated that the representational similarity between parallel non-English and English corpora, independent from the task, acts as a good barometer for multilingual performance. Our study extends the concept of representational alignment to the instance level and provides causal evidence that middle-layer alignment is sufficient for the reasoning required in MCQA tasks.

Boosting cross-lingual alignment Recent work has sought to enhance alignment through interventions, either by incorporating a separate cross-lingual alignment objective (Liu and Niehues, 2025) or fine-tuning (Li et al., 2024; Zhang et al., 2023), and has shown that improved alignment yields gains in downstream multilingual accuracy. Concurrent works also analyze inference time steering to improve multilingual performance (Wang et al., 2025b; Lim et al., 2025; Sundar et al., 2025). The idea is to steer the other language representations towards the shared latent space by computing a steering vector or a projection matrix. Our study is similar, but instead of steering representations effectively, we demonstrate how *'forcing'* alignment by patching the English activation in the middle layers helps the model rectify non-English failures.

7 Conclusions

This paper introduces instance-level alignment metrics to study LLM performance disparities across languages in discriminative tasks. We demonstrate that samples that generalize well between English and languages other than English (TS) exhibit a higher degree of alignment than samples that do not (TF), thereby establishing that cross-lingual alignment with English at the instance level is correlated with task success. To probe the causal role of alignment, we perform activation patching experiments on non-English failure instances. While patching English activations can correct incorrect non-English predictions, our control experiments help us identify specific middle layers that flip *only* with semantically equivalent patches. Overall, these findings suggest that cross-lingual semantic alignment in localized middle layers contributes causally to non-English correctness. These findings motivate future directions such as exploring targeted steering at inference time that leverages the model's high-fidelity English activations to enhance performance on languages other than English, while also studying potential downsides, including inflated confidence and the erasure of valuable local, non-English knowledge.

Limitations

Firstly, the study's limitations center on its scope, which is restricted to analyzing discriminative NLU tasks. This is due to the necessity of multilingual benchmarks that are parallel across languages. Secondly, as with most multilingual benchmarks, the benchmarks considered in the study were initially constructed in English and translated into other languages by humans, which could introduce translation artifacts (Artetxe et al., 2020). Thirdly, our study analyzes only the bilingual alignment of languages other than English with English, not between other languages.

Acknowledgements

We would like to thank the members of the CLIP lab at the University of Maryland for their constructive feedback and support. Specifically, we are grateful to Navita Goyal, Hiba El Oirghi, Dayeon (Zoey) Ki, Calvin Bao, Eric Bennet, Hieu Tran, and Osvaldo Quinjica for extensive discussions on the analyses. The authors also acknowledge the assistance of LLMs in developing this paper, in particular, using AI agents such as Claude Code to write code and create plots.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. 2025. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. 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.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english?

Jaden Fiotto-Kaufman, Alexander R. Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, et al. 2025. Nnsight and ndif: Democratizing access to open-weight foundation model internals.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models.

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.

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.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: a metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on*

Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15077–15107, Suzhou, China. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating afnd editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Aaron Mueller, Jannik Brinkmann, Millicent L. Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2024. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *CoRR*, abs/2408.01416.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 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.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining.

nosalgebraist. 2020. Interpreting gpt: The logit lens; accessed: 2025-09-25.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. 2025. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2375–2401, Vienna, Austria. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025a. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025b. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

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.

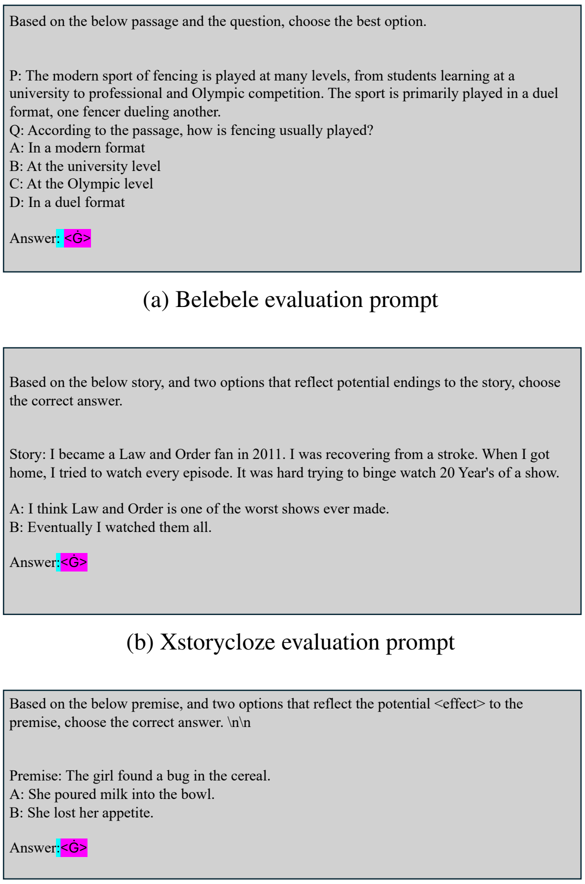
Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. 2025. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *The Thirteenth International Conference on Learning Representations*.

Bryan Wilie, Samuel Cahyawijaya, Junxian He, and Pascale Fung. 2025. High-dimensional interlingual representations of large language models. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 122–155, Vienna, Austria. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

A Appendix



Based on the below passage and the question, choose the best option.

P: The modern sport of fencing is played at many levels, from students learning at a university to professional and Olympic competition. The sport is primarily played in a duel format, one fencer dueling another.
Q: According to the passage, how is fencing usually played?
A: In a modern format
B: At the university level
C: At the Olympic level
D: In a duel format

Answer: <Ġ>

(a) Belebele evaluation prompt

Based on the below story, and two options that reflect potential endings to the story, choose the correct answer.

Story: I became a Law and Order fan in 2011. I was recovering from a stroke. When I got home, I tried to watch every episode. It was hard trying to binge watch 20 Year's of a show.

A: I think Law and Order is one of the worst shows ever made.
B: Eventually I watched them all.

Answer:<Ġ>

(b) Xstorycloze evaluation prompt

Based on the below premise, and two options that reflect the potential <effect> to the premise, choose the correct answer. \n\n

Premise: The girl found a bug in the cereal.
A: She poured milk into the bowl.
B: She lost her appetite.

Answer:<Ġ>

(c) Xcopa evaluation prompt

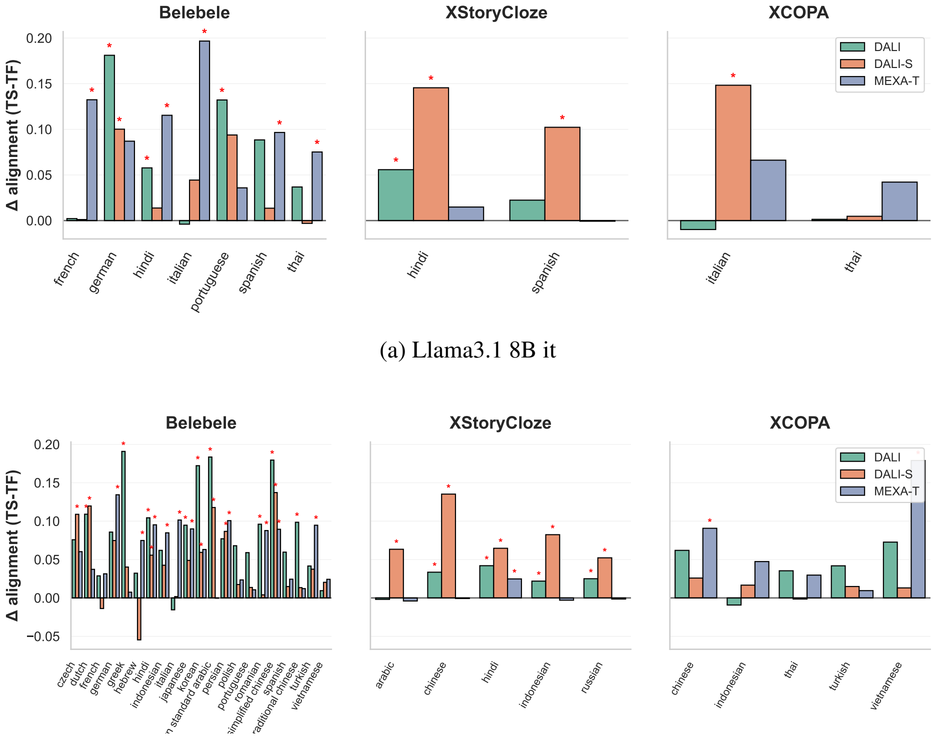
Figure 10: Evaluation prompts for different benchmarks. : refers to the penultimate token and Ġ refers to the last token respectively.

Model	Belebele	Xstorycloze	Xcopa
Llama3.1 8B, Llama3.1 8B it	Spanish, Italian, French, German, Thai, Hindi, Portuguese	English, Spanish, Hindi	Italian, Thai
Aya23 8B	Modern Standard Arabic, Simplified Chinese, Traditional Chinese, Czech, Dutch, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Turkish, Ukrainian, Vietnamese	Arabic, Spanish, Hindi, Indonesian, Russian, Chinese	Indonesian, Thai, Turkish, Vietnamese, Chinese

Table 3: Models and languages used in the cross-lingual alignment experiments

		Spanish	Italian	French	German	Thai	Portuguese	Hindi
Belebele (n=900)	$(n_{\text{TS}}, n_{\text{TF}})$	(588, 135)	(575,148)	(608,115)	(586,137)	(482, 241)	(614, 109)	(466,257)
	ΔDALI(p-val), λmax	7.5% (0.06), 14	5.9% (0.10), 9	5.1% (0.15), 14	6.8% (0.08), 14	3.9% (0.14), 14	0.3% (0.48), 13	8.3% (0.0), 14
	ΔDALI$_{st}$(p-val), λmax	2.6% (0.28), 8	6.7% (0.06), 9	6.6% (0.08), 9	4.3% (0.15), 8	1.9% (0.15), 8	-0.8% (0.56), 9	0.4% (0.4), 5
	ΔMEXA$_T$(p-val), λ_{max}	6.5% (0.08), 12	5% (0.14), 12	11.9% (0.01), 12	14.4% (0), 7	2.5% (0.09), 13	7.2% (0.06), 12	6.8% (0.02), 8
Xstorycloze (n=1511)	$(n_{\text{TS}}, n_{\text{TF}})$	(1365, 86)	.	.	.	.	.	(1240, 211)
	ΔDALI(p-val), λ_{max}	3.8% (0.01), 14	.	.	.	.	.	3.3% (0.04), 14
	ΔDALI$_{st}$(p-val), λ_{max}	0.0% (0.56), 13	.	.	.	.	.	9.5% (0.0), 13
	ΔMEXA$_T$(p-val), λ_{max}	0.0% (0.60), 7*	.	.	.	.	.	2.0% (0.01), 16
Xcopa (n=500)	$(n_{\text{TS}}, n_{\text{TF}})$	.	(374,155)	.	.	.	(262,94)	.
	ΔDALI(p-val), λ_{max}	.	2.1% (0.36), 8	.	.	.	3.9% (0.25), 16	.
	ΔDALI$_{st}$(p-val), λ_{max}	.	2.7% (0.28), 13	.	.	.	3.5% (0.22), 9	.
	ΔMEXA$_T$(p-val), λ_{max}	.	8.5% (0.12), 23	.	.	.	1.6% (0.30), 24	.

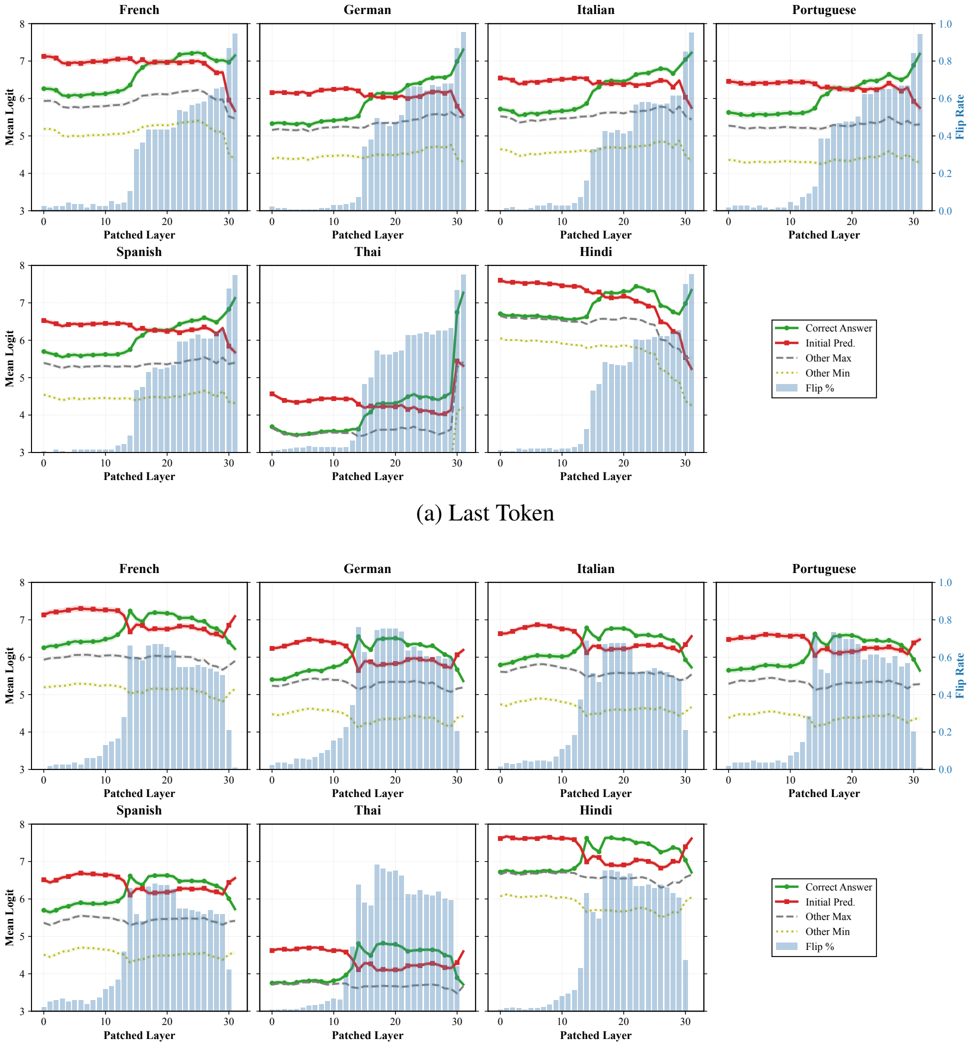
Table 4: Δ Alignment, p-val, n_{TS} and n_{TF}, and λ_{max} across the three NLU benchmarks in Llama3.1 8B.



(a) Llama3.1 8B it

(b) Aya23

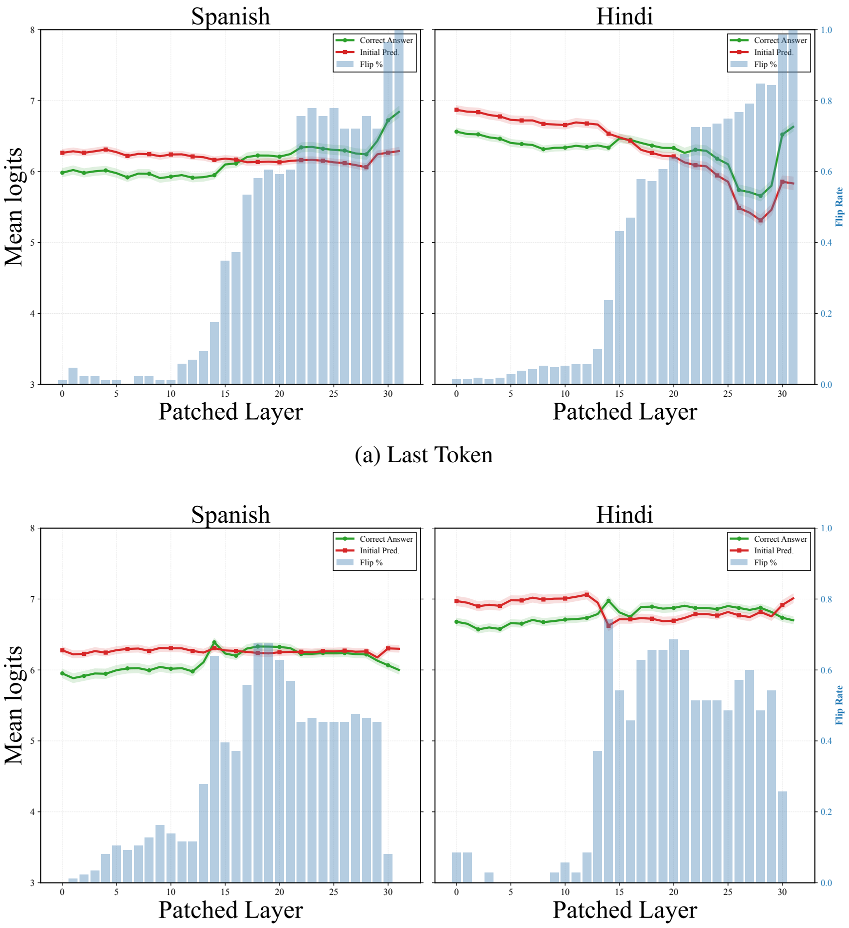
Figure 11: Δ Alignment between TS and TF samples



(a) Last Token

(b) Penultimate Token

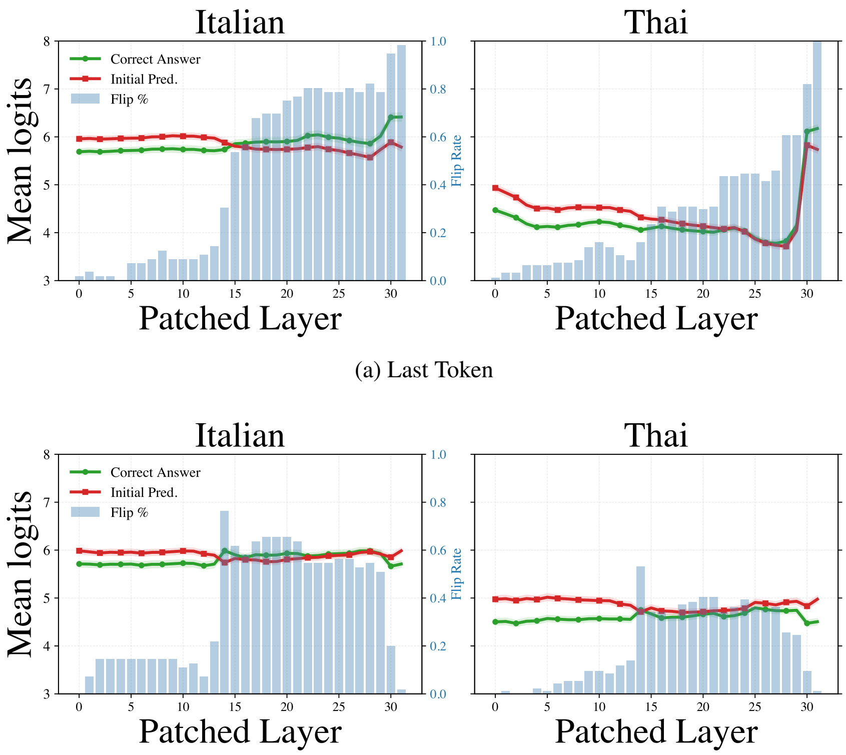
Figure 12: Patching results for the belebele benchmark across all languages in Llama 3.1 8B at last token (a) and penultimate token (b) respectively



(a) Last Token

(b) Penultimate Token

Figure 13: Patching results for the xstorycloze benchmark across all languages in Llama 3.1 8B at last token (a) and penultimate token (b) respectively



(a) Last Token

(b) Penultimate Token

Figure 14: Patching results for the xcopa benchmark across all languages in Llama 3.1 8B at last token (a) and penultimate token (b) respectively

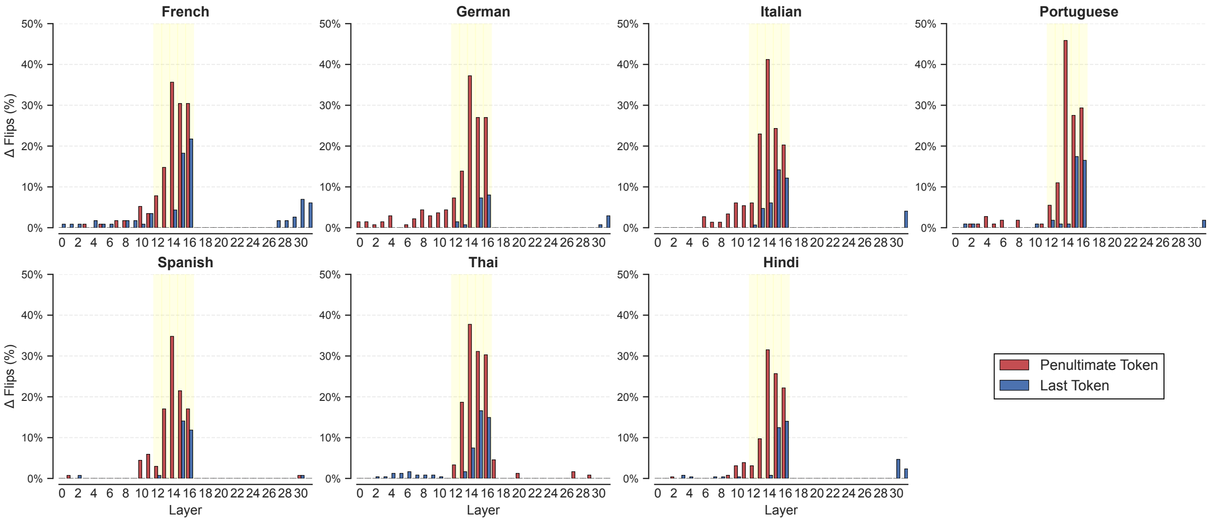


Figure 15: Control patching results for the belebele benchmark

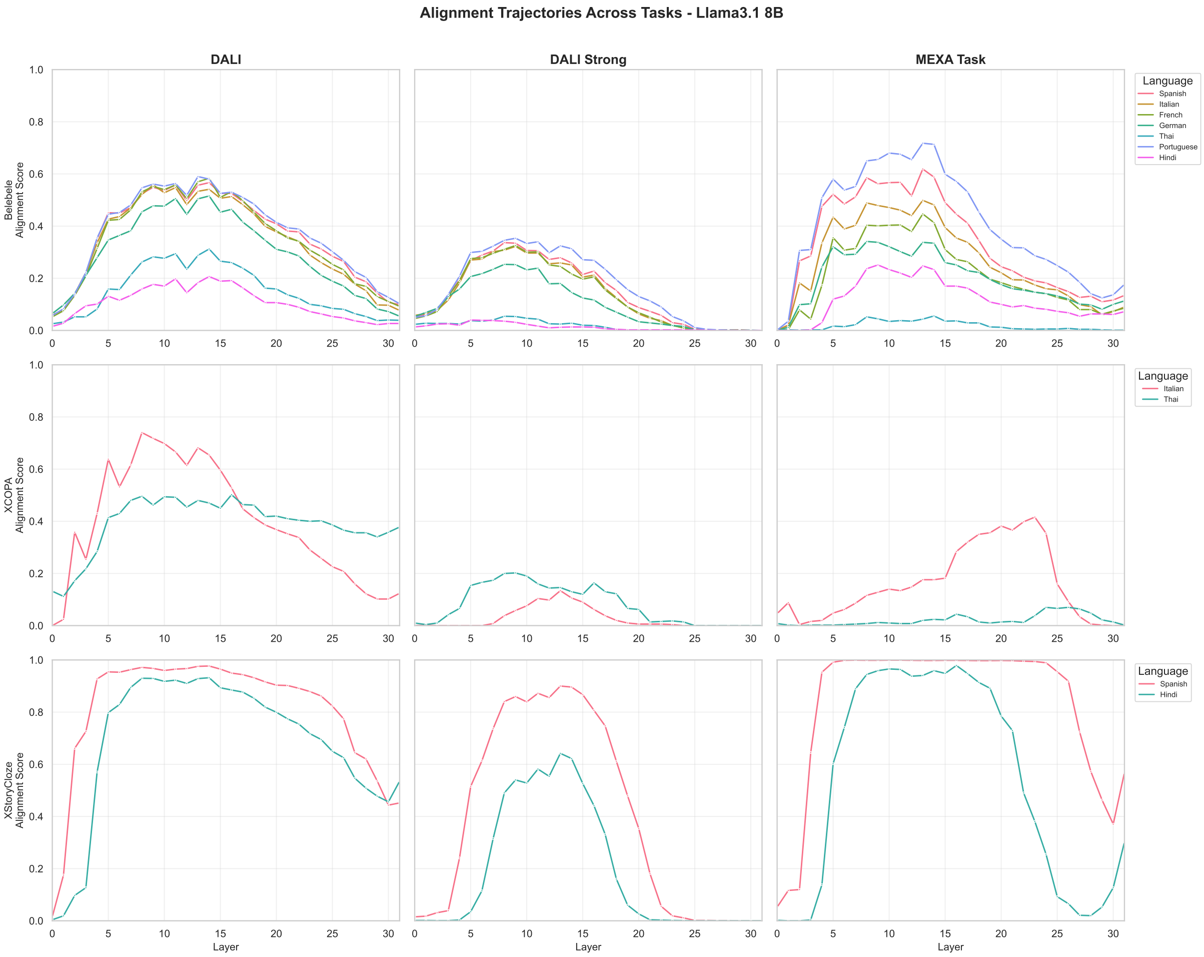


Figure 16: Alignment trajectories DALI, DALI$_{st}$, MEXA$_T$ by languages in Llama3.1-8B