On the Entity-Level Alignment in Crosslingual Consistency

Yihong Liu[1,2,*], **Mingyang Wang**[1,2,*], **François Yvon**[3,†], **and Hinrich Schütze**[1,2,†]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]Sorbonne Université, CNRS, ISIR, France
{yihong, mingyang}@cis.lmu.de

Abstract

Multilingual large language models (LLMs) are expected to recall factual knowledge consistently across languages. However, the factors that give rise to such *crosslingual consistency* – and its frequent failure – remain poorly understood. In this work, we hypothesize that these inconsistencies may arise from failures in *entity alignment*, the process of mapping subject and object entities into a shared conceptual space across languages. To test this, we assess alignment through entity-level (subject and object) translation tasks, and find that consistency is strongly correlated with alignment across all studied models, with misalignment of subjects or objects frequently resulting in inconsistencies. Building on this insight, we propose **SubSub** and **SubInj**, two effective methods that integrate English translations of subjects into prompts across languages, leading to substantial gains in both factual recall accuracy and consistency. Finally, our mechanistic analysis reveals that these interventions reinforce the entity representation alignment in the conceptual space through model's internal pivot-language processing, offering effective and practical strategies for improving multilingual factual prediction.

1 Introduction

LLMs have demonstrated remarkable capabilities in multilingual tasks, including translation, question answering, and factual recall (Petroni et al., 2019; Jiang et al., 2020; NLLB Team, 2024; Engländer et al., 2024). Among these, *crosslingual consistency* – the ability to recall the same fact correctly across different languages – has emerged as a desirable property for evaluating multilingual factual grounding (Heinzerling and Inui, 2021; Fierro and Søgaard, 2022; Qi et al., 2023). Yet, LLMs frequently exhibit *inconsistencies*, particularly when the involved languages differ in script or linguistic structure (Xing et al., 2024; Wang et al., 2025; Liu et al., 2025b). Despite growing interest in this phenomenon, the underlying mechanisms that support or limit crosslingual consistency remain poorly understood.

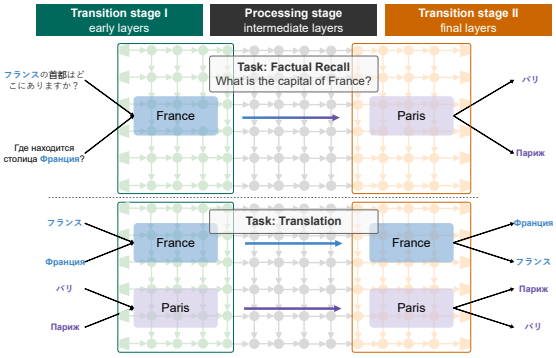


Figure 1: Analogy between consistent factual recall and entity translation. Both tasks may require mapping language-specific inputs into a shared conceptual space and projecting language-agnostic representations back into surface forms. This motivates our central hypothesis: entity alignment is important and facilitates consistent factual recall across languages.

A natural intuition suggests that consistent factual recall across languages requires a shared representation of the *entities* involved, beyond surface-level token overlap. Humans, for instance, recognize that both "フランス" (Japanese) and "Франція" (Ukrainian) refer to the same entity – *France* – and then retrieve the relevant knowledge and express it correctly in either language. This requires a form of mental *entity alignment* across languages. We hypothesize that a similar mechanism often underpins crosslingual consistency in LLMs: consistent factual recall is facilitated when entities (subject and object) are aligned across lan-

[*]Equal contribution.
[†]Equal advising.

guages in the model's internal shared language-agnostic conceptual space (see illustration of the conceptual model in Figure 1).

Such alignment facilitates the two crucial *transition stages* that occur in multilingual factual recall according to the conceptual model: (1) mapping language-specific inputs into a shared conceptual space, and (2) projecting the shared latent representation back into the correct surface realization in the target languages. Inconsistent factual recall across two languages may arise if entity alignment fails at either of these transition stages.[1] This hypothesis aligns with recent work suggesting that multilingual LLMs, especially those trained predominantly on English, often operate in an *English-centric latent space* (Wendler et al., 2024; Schut et al., 2025; Dumas et al., 2025). During multilingual factual recall, the model implicitly translates input from a non-English language into English-like conceptual representations, processes information in this *pivot-language* conceptual space, and then generates output in the desired target language (Wang et al., 2025; Lu et al., 2025). Such a pipeline inherently relies on robust crosslingual alignment, particularly for subject and object entities that serve as anchors for factual retrieval.

To operationalize this hypothesis, we introduce an *entity-level translation task* designed to probe how well models align subject and object entities between language pairs. This task instantiates the same transition stages posited by the conceptual model (cf. Figure 1), offering a concrete lens through which to assess the model's crosslingual entity-level alignment capabilities. We then evaluate the factual recall and entity-level alignment of a spectrum of LLMs from several model families across multiple sizes, using KLAR (Wang et al., 2025), a multilingual factual knowledge dataset covering 17 languages.

Our key contributions are as follows:

(i) We provide the first systematic analysis of the relationship between **crosslingual consistency** and **entity-level alignment**, using a dedicated entity translation probing task to measure entity alignment quality. **(ii)** In §4, we show that consistency is **strongly correlated** with entity alignment, and that consistent factual recall rarely occurs without correctly aligning either the subject or the object entities between two languages. **(iii)** In §5, we introduce two prompting-based remedies, **SUBSUB** and **SUBINJ**, which incorporate English translation of subjects into prompts. These methods consistently and substantially enhance recall and consistency, particularly in English-centric models. **(iv)** Finally, in §6, we analyze SUBSUB and SUBINJ through the lens of mechanistic interpretability, showing that these interventions enhance entity representation alignment through pivot-language processing and thereby facilitate consistent factual recall.

[1] Inconsistency can also stem from failure in the processing stage – i.e., the model lacks the factual knowledge so it cannot retrieve the correct object in the conceptual space, which is beyond the scope of discussion in this paper.

2 Related Work

2.1 Crosslingual Alignment

Prior work on crosslingual alignment has largely focused on *representational alignment*, where semantically similar units across languages are mapped to nearby vectors in embedding space (Artetxe and Schwenk, 2019; Reimers and Gurevych, 2020). This alignment is typically evaluated via retrieval tasks such as sentence alignment or word-level correspondence, which operationalize the notion of weak alignment (Roy et al., 2020; Hämmerl et al., 2024; Liu et al., 2025a). However, these embedding-based evaluations remain indirect and often correlate loosely with performance on downstream tasks, particularly in generation settings (Kargaran et al., 2025), where alignment must manifest at the level of surface forms. To address this gap, we evaluate alignment explicitly through input-output behavior, using translation accuracy as a proxy. This surface-level alignment provides a more functional view of whether subject and object entities are correctly realized across languages, and we show it to be strongly correlated with crosslingual consistency.

2.2 Factual Recall and Consistency

Pretrained language models have been shown to function as knowledge bases, as first proposed by Petroni et al. (2019). Following this idea, a series of studies have explored the factual knowledge stored in these models using knowledge probing techniques, focusing either on English (Roberts et al., 2020; Peng et al., 2022) or on multilingual factual knowledge (Jiang et al., 2020; Kassner et al., 2021; Yin et al., 2022; Fierro et al., 2025). Building on multilingual probing, recent work examines crosslingual consistency, the ex-

tent to which models return consistent answers to equivalent queries in different languages. Qi et al. (2023) show that LLMs often produce divergent answers across languages, highlighting inconsistencies in multilingual factual recall. Wang et al. (2025) and Lu et al. (2025) further investigate how internal representations lead to these inconsistencies, while other work traces how crosslingual consistency emerges and evolves during pretraining (Liu et al., 2025b).

In this work, we extend this line of research by identifying entity alignment – the model's ability to represent subject and object entities consistently across languages – as a key factor underlying crosslingual consistency. We show a strong correlation between entity alignment and crosslingual consistency, and consistent factual recall hardly occurs when the model fails to align entities. Building on this insight, we propose two prompting-based interventions, which enhance entity alignment and thereby facilitate crosslingual factual consistency. Finally, we perform a mechanistic interpretability analysis to explain the success of these interventions, offering both theoretical insights and practical strategies for improving multilingual factual prediction in LLMs.

3 Methodology

3.1 Languages, Models, and Dataset

Languages. We consider 17 languages that span 8 language families and use 8 different scripts: Arabic (**ara_Arab**), Catalan (**cat_Latn**), Chinese (**zho_Hans**), Dutch (**nld_Latn**), English (**eng_Latn**), French (**fra_Latn**), Greek (**ell_Grek**), Hebrew (**heb_Hebr**), Hungarian (**hun_Latn**), Japanese (**jpn_Jpan**), Korean (**kor_Kore**), Persian (**fas_Arab**), Russian (**rus_Cyrl**), Spanish (**spa_Latn**), Turkish (**tur_Latn**), Ukrainian (**ukr_Cyrl**), and Vietnamese (**vie_Latn**).

Models. We evaluate a diverse suite of 12 decoder-only language models spanning 4 major model families: **LLaMA** (Grattafiori et al., 2024), **Qwen** (Yang et al., 2025), **Gemma** (Gemma Team et al., 2025), and **OLMo** (Team OLMo et al., 2025). The first three are trained on highly multilingual corpora, while **OLMo** is primarily trained on English data. From the LLaMA family, we consider `Llama-3.2-1B`, `Llama-3.2-3B`, and `Llama-3.1-8B`. From the Qwen family, we consider `Qwen3-1.7B-Base`, `Qwen3-4B-Base`, and `Qwen3-8B-Base`. From the Gemma family, we consider `gemma-3-1b-pt`, `gemma-3-4b-pt`, and `gemma-3-12b-pt`. Finally, we consider `OLMo-2-0425-1B`, `OLMo-2-1124-7B`, and `OLMo-2-1124-13B` from the OLMo series. This selection allows us to systematically study model behavior across size and family.

Multilingual Factual Dataset. We conduct our investigation using KLAR (Wang et al., 2025), a multilingual factual knowledge probing dataset. Our evaluation covers **2,619** language-agnostic facts spanning **20** distinct relation types (cf. Table 3 in §A). Each fact is represented as a triple (s_i, r_i, o_i), where s_i and o_i are the subject and object entities, and r_i is the relation. We denote the full language-agnostic set of facts as $\mathcal{F} = \{(s_i, r_i, o_i)\}_{i=1}^{N}$. For each fact $(s_i, r_i, o_i) \in \mathcal{F}$ and each language l, KLAR provides corresponding language-specific realizations (s_i^l, r_i, o_i^l), where s_i^l and o_i^l are the subject and object translated into language l. Thus, for any two languages l and l', the instances (s_i^l, r_i, o_i^l) and $(s_i^{l'}, r_i, o_i^{l'})$ represent aligned expressions of the same underlying fact in the two languages. KLAR also provides a set of language-specific prompt templates for each relation, which we use to construct factual recall queries in different languages (cf. §3.3).

3.2 Entity-Level Alignment Evaluation

Task Formulation. To evaluate crosslingual entity-level alignment, we formulate a multilingual entity translation task over the language-agnostic fact set $\mathcal{F} = \{(s_i, r_i, o_i)\}_{i=1}^{N}$. For each language pair (l_1, l_2) and each fact $(s_i, r_i, o_i) \in \mathcal{F}$, we construct two translation sub-tasks: (1) translating the subject entity $s_i^{l_1} \rightarrow s_i^{l_2}$, and (2) translating the object entity $o_i^{l_1} \rightarrow o_i^{l_2}$. The model is prompted in the source language l_1 and is expected to generate the corresponding entity in the target language l_2. Each prompt includes a 3-shot demonstration consisting of aligned entity pairs drawn from other facts in the dataset. An example of a subject translation prompt from Japanese to English is displayed below:

日本語: フランス - English: France
日本語: セルビア - English: Serbia
日本語: イタリア - English: Italy
日本語: イギリス - English:

The model is expected to complete the last translation with the correct entity name in the target language ("*United Kingdom*" in this case). **Metrics.** For each direction $(l_1 \rightarrow l_2)$ and each fact $(s_i, r_i, o_i) \in \mathcal{F}$, we evaluate whether the model's generated output contains the correct target entity. Specifically, we define:

Subject translation accuracy:

$$\mathrm{ACC}^{\mathrm{sub}}_{l_1 \rightarrow l_2} = \frac{1}{|\mathcal{F}|} \sum_{i=1}^{|\mathcal{F}|} \mathbf{1}\left[s_i^{l_2} \subseteq \mathcal{M}(s_i^{l_1})\right]$$

Object translation accuracy:

$$\mathrm{ACC}^{\mathrm{obj}}_{l_1 \rightarrow l_2} = \frac{1}{|\mathcal{F}|} \sum_{i=1}^{|\mathcal{F}|} \mathbf{1}\left[o_i^{l_2} \subseteq \mathcal{M}(o_i^{l_1})\right]$$

Joint translation accuracy:

$$\mathrm{ACC}^{\mathrm{both}}_{l_1 \rightarrow l_2} = \frac{1}{|\mathcal{F}|} \sum_{i=1}^{|\mathcal{F}|} \mathbf{1}\Big[s_i^{l_2} \subseteq \mathcal{M}(s_i^{l_1}) \wedge o_i^{l_2} \subseteq \mathcal{M}(o_i^{l_1})\Big]$$

where $\mathcal{M}(x)$ denotes the model's *full generated output* when prompted with x, and $\mathbf{1}[\cdot]$ is the indicator function. Our evaluation differs from prior work that only checks the first generated token (Geva et al., 2023; Qi et al., 2023; Hernandez et al., 2024), ensuring correctness at the string level, especially for *multi-token entities*. To assess alignment between l_1 and l_2, we average the translation accuracies in both directions. Specifically:

$$\mathrm{Align}^{\mathrm{sub}}(l_1, l_2) = \frac{\mathrm{ACC}^{\mathrm{sub}}_{l_1 \rightarrow l_2} + \mathrm{ACC}^{\mathrm{sub}}_{l_2 \rightarrow l_1}}{2}$$

$$\mathrm{Align}^{\mathrm{obj}}(l_1, l_2) = \frac{\mathrm{ACC}^{\mathrm{obj}}_{l_1 \rightarrow l_2} + \mathrm{ACC}^{\mathrm{obj}}_{l_2 \rightarrow l_1}}{2}$$

$$\mathrm{Align}^{\mathrm{both}}(l_1, l_2) = \frac{\mathrm{ACC}^{\mathrm{both}}_{l_1 \rightarrow l_2} + \mathrm{ACC}^{\mathrm{both}}_{l_2 \rightarrow l_1}}{2}$$

The alignment scores are computed across all $\binom{17}{2} = 136$ unique language pairs, yielding a comprehensive view of entity-level alignment.

3.3 Factual Recall Evaluation

Task Formulation. Given the multilingual fact set $\mathcal{F} = \{(s_i, r_i, o_i)\}_{i=1}^{N}$, KLAR provides a language-specific prompt template for each relation r_i in every language l. Each fact can be instantiated in language l as a *query-answer pair* (q_i^l, o_i^l), where q_i^l is the prompt formed using s_i^l and r_i, and o_i^l is the language-specific realization of the object o_i. For example, the fact (*France, capital, Paris*) may be rendered in English as "*Where is France's capital located? The answer is:*" with the expected answer "*Paris*". This setup ensures that for each fact, we can construct 17 language-specific queries, one per language. The queries $q_i^{l_1}$ and $q_i^{l_2}$ for languages l_1 and l_2 share the same underlying fact and differ only in surface realization. To facilitate factual recall, we apply a 3-shot prompting strategy analogous to that used in the entity translation task (cf. §3.2). Each prompt is preceded by three example question-answer pairs in the same language and relation type, selected from other facts in the dataset. This few-shot design provides a contextual signal for relation-specific generation while preserving the language setting.

Metrics. We use **accuracy** and **crosslingual consistency** to evaluate model performance. The per-language factual recall accuracy is defined as:

$$\mathrm{ACC}(l) = \frac{1}{|\mathcal{F}|} \sum_{i=1}^{|\mathcal{F}|} \mathbf{1}\left[o_i^l \subseteq \mathcal{M}(q_i^l)\right]$$

where $\mathcal{M}(q_i^l)$ denotes the model's complete generation for prompt q_i^l, and correctness is judged by string inclusion as described in §3.2. To assess crosslingual consistency of a language pair (l_1, l_2), we compute the Jaccard index:[2]

$$\mathrm{CO}(l_1, l_2) = \frac{\sum_{i=1}^{|\mathcal{F}|} \mathbf{1}[o_i^{l_1} \subseteq \mathcal{M}(q_i^{l_1}) \wedge o_i^{l_2} \subseteq \mathcal{M}(q_i^{l_2})]}{\sum_{i=1}^{|\mathcal{F}|} \mathbf{1}[o_i^{l_1} \subseteq \mathcal{M}(q_i^{l_1}) \vee o_i^{l_2} \subseteq \mathcal{M}(q_i^{l_2})]}$$

We compute $\mathrm{CO}(l_1, l_2)$ for all $\binom{17}{2} = 136$ language pairs to obtain a comprehensive map of crosslingual factual consistency.

4 Relationship Between Crosslingual Consistency and Entity Alignment

4.1 Consistency Correlated with Alignment

We examine whether alignment is *correlated* with consistency across language pairs in practice. Specifically, we perform a correlation analysis between consistency $\mathrm{CO}(l_1, l_2)$ and each of the alignment metrics: $\mathrm{Align}^{\mathrm{sub}}(l_1, l_2)$, $\mathrm{Align}^{\mathrm{obj}}(l_1, l_2)$, and $\mathrm{Align}^{\mathrm{both}}(l_1, l_2)$. For each model, we compute Pearson correlation coefficients over the 136 language pairs (cf. Figure 2).

[2] We adopt the definition of consistency used in prior work (Jiang et al., 2020; Wang et al., 2025), which is stricter than mere agreement: it requires that the model's predictions be both identical *and* correct across the two languages.

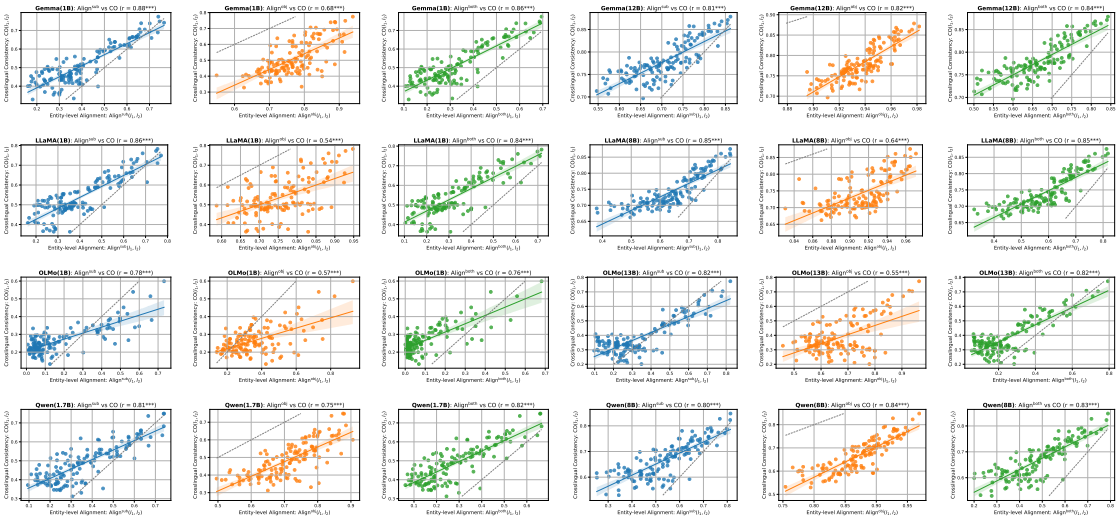


Figure 2: Correlation between entity-level alignment and crosslingual consistency. Each subplot displays the relationship between one alignment metric, i.e., $\text{Align}^{\text{sub}}(l_1, l_2)$, $\text{Align}^{\text{obj}}(l_1, l_2)$, or $\text{Align}^{\text{both}}(l_1, l_2)$, and crosslingual consistency $\text{CO}(l_1, l_2)$ for a given model. Each point represents a language pair. The gray dashed line (the diagonal $y = x$) separates the region into one half where consistency is higher (above the line) and one half where alignment is higher (below the line). Strong and statistically significant correlations are observed across models, supporting our hypothesis that alignment is highly associated with consistency.

We observe strong and statistically significant correlations between consistency and alignment across models. This result supports our hypothesis that crosslingual consistency is tightly linked to the model's ability to align entities across languages. In all cases, the Pearson correlation is quite high, exceeding 0.7 for subject/both alignment and 0.5 for object alignment, significant at $p < 0.001$. Notably, while high alignment does not always guarantee high consistency – particularly in the case of object alignment (e.g., LLaMA (1B)) – we rarely observe the opposite: high consistency with low alignment score. This asymmetry suggests that alignment is closely associated with, and may be a prerequisite for, crosslingual consistency, though it is not by itself sufficient – a pattern we examine further in §4.2.

Among the three alignment metrics, subject and joint alignment correlate more strongly with consistency than object alignment. This asymmetry may reflect the task structure of factual recall: the subject entity is always present in the prompt and serves as the anchor for factual prediction. If the model fails to align the subject of a fact across two languages in the first place, it is less likely to consistently recall facts containing the subject. Therefore, subject alignment may play an important role in facilitating consistency. Based on this insight, we present two simple remedies for improving the factual recall in §5.

OLMo exhibits weaker correlation, particularly with object-level alignment. Although OLMo still shows statistically significant correlations, they are generally weaker than those of multilingual models like Qwen or LLaMA. A likely cause is its limited exposure to factual content in non-English languages during pretraining (Liu et al., 2025b), which impairs both factual recall and entity alignment. Additionally, English-centric models often rely on English as a pivot language (Wendler et al., 2024) in the language-agnostic conceptual space, but the prompts do not provide any explicit English signal, weakening the model's ability to bridge language-specific inputs and *correct* concepts. As shown in §5, injecting English subjects into prompts can mitigate this limitation and substantially improve factual recall.

Consistency is bounded by object alignment. We observe (Figure 2) that crosslingual consistency is almost always lower than object alignment scores across language pairs. This is made explicit in Figure 3, which overlays consistency against object alignment scores for all language pairs across models. The vast majority of points lie below the diagonal, indicating that a model may not consistently recall a fact across two lan-

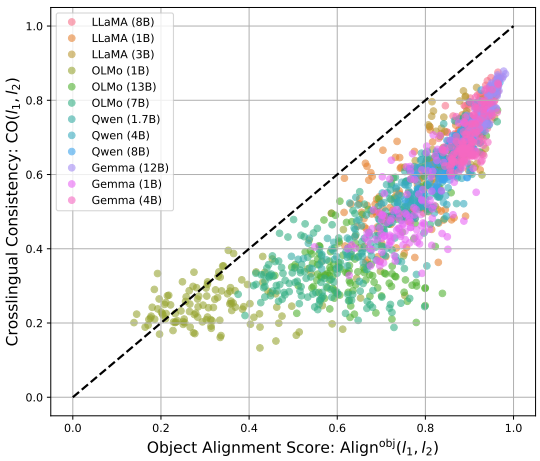


Figure 3: The boundedness relationship between consistency $\mathrm{CO}(l_1, l_2)$ and object alignment score $\mathrm{Align}^{\mathrm{obj}}(l_1, l_2)$. Each point indicates a language pair, while different colors indicate different models. $\mathrm{CO}(l_1, l_2)$ is almost always upper-bounded by $\mathrm{Align}^{\mathrm{obj}}(l_1, l_2)$ except for OLMo (1B).

guages if it fails to align the object entity. This pattern likely arises because both consistent factual recall and entity translation require the model to generate the correct object across languages. The only exception is OLMo (1B), possibly due to its English-centric training data and weaker multilingual grounding, as discussed earlier.

Overall, these findings reinforce our core claim that good entity (subject and object) alignment facilitates crosslingual consistency.

4.2 Consistent Facts Are Entity-Aligned

We have shown that entity alignment and crosslingual consistency are strongly correlated at the aggregate level. In this section, we aim to deepen our understanding of their relationship by analyzing individual fact instances. Specifically, we ask two questions: (1) *To what extent does entity alignment support crosslingual consistency?*, and (2) *Can non-aligned facts still be consistently recalled across languages?* To answer these questions, we examine each fact and determine whether it is crosslingually consistent or entity-aligned.[3] This fine-grained analysis allows us to characterize the role of entity alignment in enabling consistency, and to assess how often consistency emerges even in the absence of entity alignment.

We say that a fact f_i is crosslingually ***consistent*** for languages l_1 and l_2 if the model generates outputs that contain the expected answers in both languages. Formally, this mirrors the definition used in our consistency metric (cf. §3.3):

$$\mathcal{C}(l_1, l_2) = \left\{ f_i \;\middle|\; o_i^{l_1} \subseteq \mathcal{M}(q_i^{l_1}) \;\wedge\; o_i^{l_2} \subseteq \mathcal{M}(q_i^{l_2}) \right\}$$

Similarly, we define crosslingually ***inconsistent*** facts for language l_1 and l_2 as the complement of $\mathcal{C}(l_1, l_2)$, which is referred to as $\mathcal{I}(l_1, l_2)$.

We further categorize each fact f_i based on whether its subject or object entities are aligned between two languages, i.e., whether the entities can be correctly translated between l_1 and l_2. We define the set of ***aligned*** facts as:

$$\mathcal{A}(l_1, l_2) = \{ f_i \mid (s_i^{l_2} \subseteq \mathcal{M}(s_i^{l_1}) \vee s_i^{l_1} \subseteq \mathcal{M}(s_i^{l_2})) \vee (o_i^{l_2} \subseteq \mathcal{M}(o_i^{l_1}) \vee o_i^{l_1} \subseteq \mathcal{M}(o_i^{l_2})) \}$$

That is, a fact is considered aligned if at least one of the four translation tasks succeeds.[4] This relaxed formulation reflects two intuitions: (1) correct alignment of either the subject or object may already provide a strong conceptual anchor to access the underlying fact and (2) asymmetric translations (e.g., correct for $l_1 \rightarrow l_2$ but not for $l_2 \rightarrow l_1$) are common, especially when involving low-resource languages. The set of ***non-aligned*** facts $\mathcal{N}(l_1, l_2)$ is then the complement of $\mathcal{A}(l_1, l_2)$.

We report the proportion of consistent facts $\mathcal{C}(l_1, l_2)$, inconsistent facts $\mathcal{I}(l_1, l_2)$, aligned facts $\mathcal{A}(l_1, l_2)$, and non-aligned facts $\mathcal{N}(l_1, l_2)$ for each

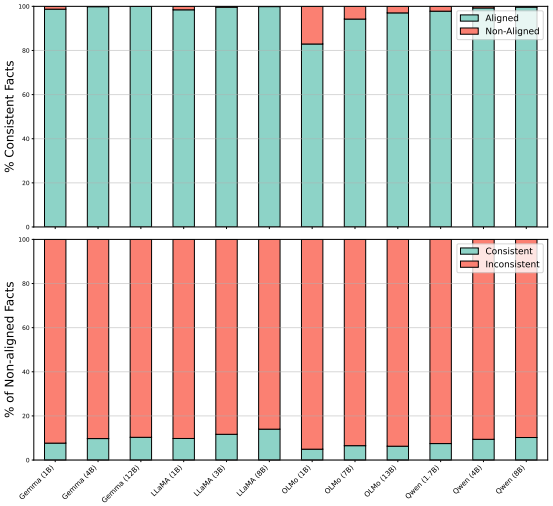


Figure 4: Entity alignment and consistency analysis across models. Most consistent facts are entity-aligned, and misalignment leads to inconsistency.

[3] Throughout this section, for simplicity, a fact refers to $(\{l_1, l_2\}, f_i)$, i.e., a fact with respect to a given language pair.

[4] Here we use *alignment* at the *instance level*, i.e., whether the subject or object entity of a given fact can be correctly translated between l_1 and l_2. This differs from the usage in §4.1, where alignment refers to *aggregated alignment scores* over translation performance across all facts.

	Gemma (1B)	Gemma (4B)	Gemma (12B)	LLaMA (1B)	LLaMA (3B)	LLaMA (8B)	OLMo (1B)	OLMo (7B)	OLMo (13B)	Qwen (1.7B)	Qwen (4B)	Qwen (8B)
$\mathcal{C}(l_1, l_2)$	124K (0.35)	194K (0.54)	229K (0.64)	134K (0.38)	188K (0.53)	215K (0.60)	42K (0.12)	72K (0.20)	87K (0.24)	113K (0.32)	159K (0.45)	186K (0.52)
$\mathcal{I}(l_1, l_2)$	231K (0.65)	162K (0.46)	127K (0.36)	222K (0.62)	169K (0.47)	142K (0.40)	314K (0.88)	284K (0.80)	269K (0.76)	243K (0.68)	197K (0.55)	171K (0.48)
$\mathcal{A}(l_1, l_2)$	335K (0.94)	353K (0.99)	355K (1.00)	335K (0.94)	350K (0.98)	354K (0.99)	211K (0.59)	293K (0.82)	315K (0.88)	323K (0.91)	344K (0.97)	350K (0.98)
$\mathcal{N}(l_1, l_2)$	21K (0.06)	3K (0.01)	1K (0.00)	22K (0.06)	6K (0.02)	2K (0.01)	145K (0.41)	64K (0.18)	41K (0.12)	33K (0.09)	12K (0.03)	6K (0.02)

Table 1: Proportion of consistent ($\mathcal{C}(l_1, l_2)$), inconsistent ($\mathcal{I}(l_1, l_2)$), aligned ($\mathcal{A}(l_1, l_2)$), and non-aligned crosslingual facts ($\mathcal{N}(l_1, l_2)$) for each model. Each fact in 2,619 factual queries is evaluated over all $\binom{17}{2} = 136$ language pairs, resulting in a total of $2619 \times 136 = 356{,}184$ evaluated facts per model.

model in Table 1. Figure 4 provides a more detailed breakdown: (1) among all consistent facts $\mathcal{C}(l_1, l_2)$, what fraction are aligned ($\mathcal{A}(l_1, l_2)$) vs. non-aligned ($\mathcal{N}(l_1, l_2)$), and (2) among all non-aligned facts $\mathcal{N}(l_1, l_2)$, what proportion are consistent ($\mathcal{C}(l_1, l_2)$) vs. inconsistent ($\mathcal{I}(l_1, l_2)$). Together, they illustrate how much consistency can be attributed to entity alignment, and whether models can still achieve consistency without it.

Models exhibit stronger entity alignment than crosslingual consistency. As shown in Table 1, nearly all models achieve high rates of entity alignment across language pairs, often aligning over 95% of facts. For example, Gemma (12B) and LLaMA (8B) align more than 99% of all evaluated crosslingual facts. In contrast, their consistency rates remain notably lower (64% and 60%, respectively). This discrepancy suggests that entity alignment is a foundational skill models acquire more easily during pretraining, whereas achieving consistency demands additional capabilities such as relation representation and knowledge retrieval (required in the processing stage in Figure 1). With a larger model size, such capabilities are enhanced; therefore, the consistency is substantially improved (e.g., from 0.35 to 0.64 when comparing Gemma (1B) and Gemma (12B)). The OLMo family, in particular, shows weak alignment and consistency, likely due to its English-centric pretraining data.

Crosslingual consistency is tightly conditioned on entity alignment. Figure 4 shows that the vast majority of consistent facts are entity-aligned, particularly for highly multilingual models like Gemma and Qwen, where over 99% of consistent facts involve successful entity alignment (either the subject, object, or both). In contrast, consistent recall among non-aligned facts is rare: fewer than 10% of such cases achieve crosslingual consistency, regardless of model family or size.[5] This asymmetry suggests that consistency across languages does not arise from chance or superficial pattern matching, but instead critically depends on the model's ability to semantically align entities. Entity alignment, therefore, acts as a prerequisite or bottleneck for achieving consistency. These findings underscore the importance of entity-level alignment and imply that enhancing alignment – particularly for low-resource languages – may directly improve a model's ability to maintain consistent factual knowledge across linguistic boundaries.

5 Remedies: SUBSUB and SUBINJ

In §4, we show that crosslingual consistency is closely tied to a model's ability to align entities across languages. However, many LLMs, particularly English-centric ones, struggle with this alignment, reflecting weaknesses in the transition from language-specific surface forms to language-agnostic conceptual representations. This observation motivates a simple yet effective idea: if a model fails to map the subject into a shared conceptual space, why not provide it with a reliable lexical anchor? Since most LLMs are extensively trained on English and often use English as an implicit pivot language in the conceptual space (Wendler et al., 2024; Wang et al., 2025), integrating English subject information can help the model bypass the subject mapping step and thereby improve factual recall across languages.

5.1 Prompt Formulation

We propose two prompting-based remedies to implement this idea. **SUBSUB** replaces the subject in the original-language prompt with its English equivalent, enforcing direct entity alignment. This idea is similar to the substitution method proposed by Yang et al. (2024) to facilitate the multi-hop reasoning in factual recall. For example, we replace "フランス" with "France" and the final

[5]These non-aligned but consistent facts may be a form of "memorization", where the model directly memorizes the fact as a whole in the respective language instead of transferring the knowledge across languages through a shared conceptual space, as shown by Liu et al. (2025b).

Model	Recall (ACC)					Consistency (CO)				
	Base	SUBSUB	SUBINJ	↑SUBSUB (%)	↑SUBINJ (%)	Base	SUBSUB	SUBINJ	↑SUBSUB (%)	↑SUBINJ (%)
Gemma (1B)	0.51	0.56	**0.58**	9.0	13.7	0.51	0.56	**0.60**	10.7	17.3
Gemma (4B)	0.66	0.71	**0.72**	6.5	8.3	0.70	0.75	**0.77**	8.0	10.6
Gemma (12B)	0.73	0.75	**0.76**	2.8	4.1	0.78	0.81	**0.83**	4.4	6.2
LLaMA (1B)	0.53	0.60	**0.61**	11.7	14.9	0.54	0.61	**0.63**	13.2	17.8
LLaMA (3B)	0.65	0.71	**0.72**	8.6	10.7	0.68	0.75	**0.77**	11.1	14.3
LLaMA (8B)	0.71	0.75	**0.76**	5.6	7.2	0.74	0.80	**0.82**	7.7	10.0
OLMo (1B)	0.27	0.28	**0.31**	2.4	11.2	0.27	0.24	**0.28**	-8.7	4.8
OLMo (7B)	0.38	0.45	**0.47**	19.5	25.1	0.35	0.43	**0.46**	23.3	31.6
OLMo (13B)	0.43	0.54	**0.56**	25.9	31.6	0.39	0.52	**0.55**	35.3	43.9
Qwen (1.7B)	0.48	0.51	**0.54**	6.3	11.7	0.49	0.52	**0.56**	7.3	14.5
Qwen (4B)	0.59	0.63	**0.64**	5.9	8.4	0.60	0.66	**0.68**	8.7	12.4
Qwen (8B)	0.65	0.68	**0.69**	4.0	6.3	0.67	0.71	**0.73**	6.0	9.2

Table 2: Performance of SUBSUB and SUBINJ compared to Base. Both methods outperform the baseline in terms of factual recall (ACC) and crosslingual consistency (CO). SUBINJ yields stronger improvements than SUBSUB across all models. Percentages indicate relative improvements over the baseline.

prompt in Japanese would then be "Franceの首都はどこにありますか？答えは：" (translation: *Where is France's capital located? The answer is:*). **SUBINJ** appends the English subject alongside the native-language form, preserving natural phrasing while encouraging the model to align the entities (e.g., "フランス (France) の首都はどこにありますか？答えは："). The two proposed remedies align with recent ideas of leveraging *code-switching* for better comprehension of the original text (Mohamed et al., 2025).

These remedies are conceptually related to the *vector interventions* of Lu et al. (2025), who show that factual inconsistencies often arise because models fail to sufficiently engage their reliable *English-centric* factual recall pipeline. By injecting carefully constructed vectors into intermediate layers, they steer models to re-activate these latent pathways, thereby improving factual recall across languages. Our approach can be seen as a *prompting-level analogue*: instead of modifying hidden states directly, we inject an explicit English signal into the input. This encourages the model to follow a similar trajectory as vector interventions – activating English-centric recall mechanisms – while remaining lightweight and training-free.

5.2 Results and Discussion

Table 2 summarizes the recall and consistency gains of both methods across models. Additionally, Figure 5 presents a fine-grained analysis by language family and script.

Both SUBSUB and SUBINJ yield consistent improvements in factual recall and consistency. As shown in Table 2, both strategies lead to performance gains across all model families and sizes. SUBINJ consistently outperforms SUBSUB, albeit by a modest margin, indicating that appending the English subject preserves the information of the original prompt while providing a strong alignment cue. For multilingual LLMs, the improvements are most pronounced in smaller models. For example, `gemma-3-1b-pt` shows an improvement of 13.7% in ACC and 17.3% in CO under SUBINJ. However, the gains become small as the size increases, suggesting that larger models already possess stronger internal alignment capabilities. This aligns with recent findings from Lim et al. (2025) that larger models are more capable of retrieving knowledge embedded across typologically different languages.[6]

English-centric models benefit substantially from the remedies, especially for larger models. Unlike multilingual models, where gains drop with increasing model size, English-centric models, i.e., OLMo, exhibit the opposite trend: performance improvements from SUBINJ and SUBSUB become more pronounced as the model scales up. For instance, `OLMo-2-1124-13B` shows a remarkable increase of over 30% in ACC and over 40% in CO. This reverse pattern reflects a key difference in how multilingual and English-centric models process crosslingual input. While multilingual models already maintain robust entity alignment across languages, English-centric models lack such alignment and rely heavily on English as a pivot language in their latent conceptual

[6] Lim et al. (2025) found that while larger models have stronger multilingual capabilities, e.g., crosslingual language understanding, hidden states of different languages are more likely to dissociate from the shared conceptual space compared to smaller models.

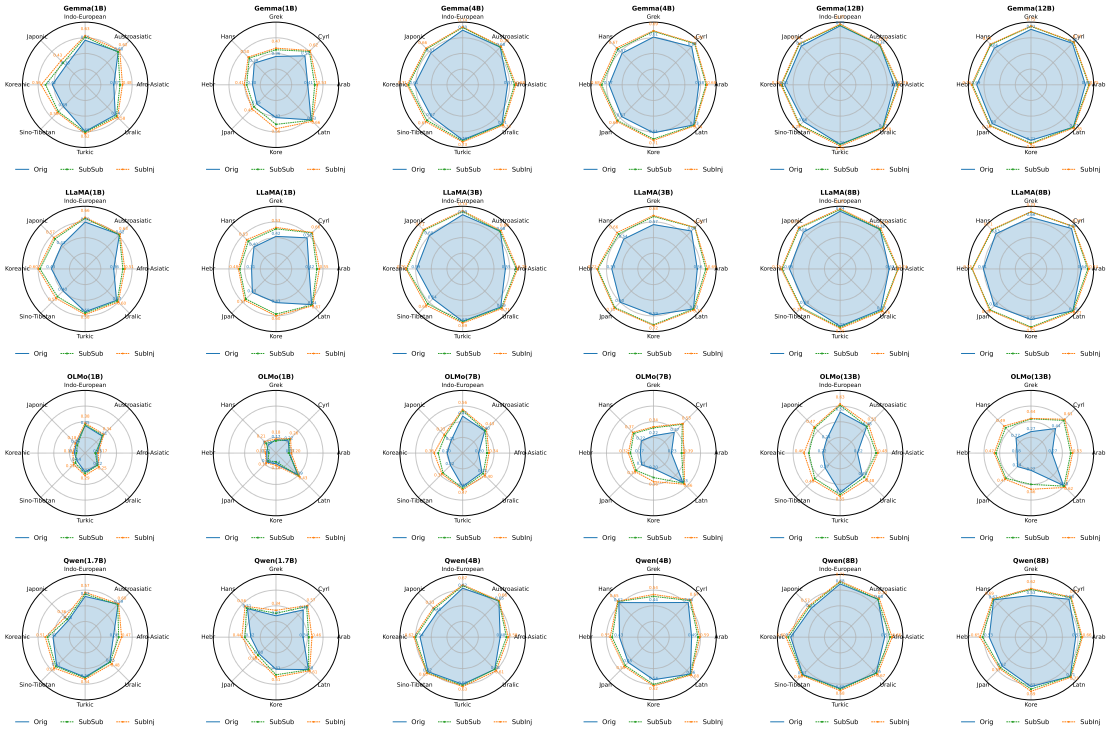


Figure 5: Radar plots comparing Baseline (Orig), SUBSUB, and SUBINJ factual recall performance (ACC) across language families and script groups. SUBINJ consistently improves recall across all categories, especially in English-centric models (i.e., OLMo model families) and in non-Latin scripts.

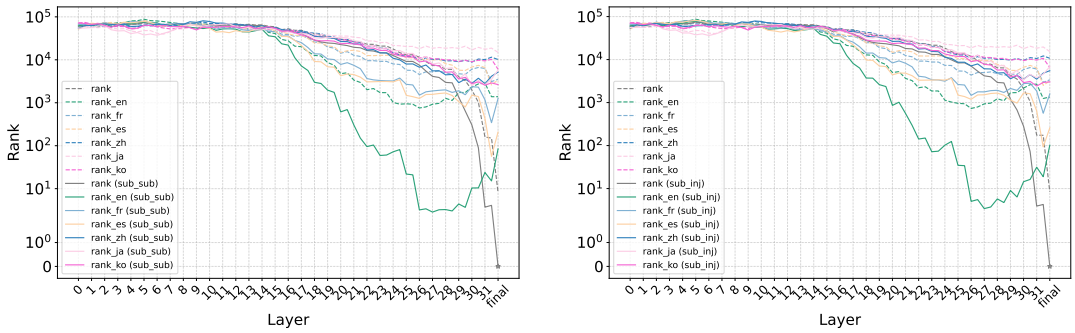
space. The primary challenge thus lies in transitioning from language-specific inputs into this shared latent space. This transition depends critically on both strong entity alignment and sufficient model capacity, explaining why the effectiveness of both methods grows with model size.

SUBSUB and SUBINJ consistently improve factual recall across language families and scripts. As illustrated in Figure 5, both methods yield consistent gains across all language families and script groups. Improvements are particularly pronounced for non-Latin scripts such as Cyrillic, Arabic, and Chinese (Hans), likely because these languages lack surface-level token overlap with English. This makes crosslingual entity alignment, especially subject alignment, more challenging. Our approach, by explicitly leveraging English subjects, encourages direct entity alignment and thereby facilitates consistency, likely by addressing cases where subject alignment is missing or suboptimal. Substantial gains are also observed in typologically diverse language families such as Semitic and Slavic, showing the robustness of our methods across linguistic boundaries. Notably, English-centric models benefit the most, with marked improvements in nearly all non-Latin scripts and language families. This underscores the effectiveness of integrating English subjects as a way to facilitate entity alignment, which further reinforces the role of English as a conceptual pivot language in multilingual factual recall.

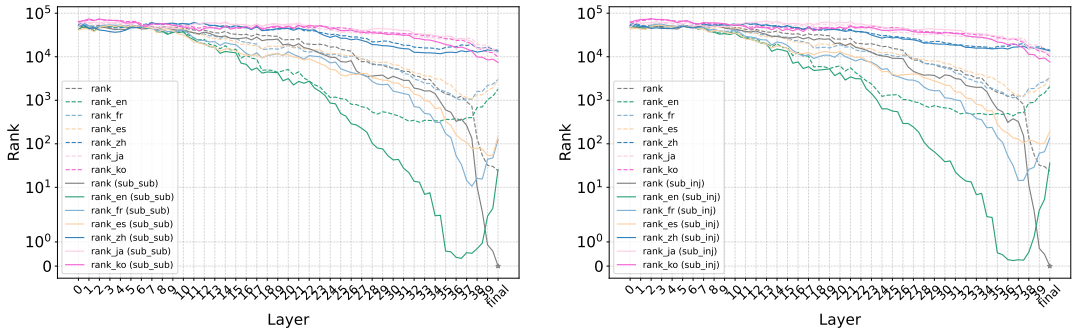
6 Mechanistic Interpretability

To better understand why SUBSUB and SUBINJ improve crosslingual factual recall, we perform a mechanistic analysis of how these interventions affect internal model representations. Specifically, we apply Logit Lens (Nostalgebraist, 2020) to project intermediate hidden states at each layer onto the output vocabulary and track the rank of the target object token across layers. A lower rank indicates that the correct answer is more readily accessible from that representation, thus reflecting stronger factual grounding.

We analyze how the concept of the target object entity is represented and decoded across languages by plotting its rank curves in the model's intermediate layers. For each prompt in a given input language, we track the rank of (i) the correct object token in that language, and (ii) its equivalents in six other languages: English (en), French (fr), Spanish (es), Chinese (zh), Japanese (ja), and Korean (ko). This allows us to examine how well the model encodes a language-agnostic representation of the object entity. Figure 6 displays the



(a) **LLaMA (8B)**: Left – with SUBSUB; Right – with SUBINJ.



(b) **OLMo (13B)**: Left – with SUBSUB; Right – with SUBINJ.

Figure 6: **Logit Lens analysis of object token ranks across model layers.** We plot the rank of the target object token (lower is better) in its input language, along with its equivalents in six other languages, using lines of different colors. Dashed lines represent the original prompts, while solid lines correspond to prompts modified with SUBSUB and SUBINJ. Both interventions consistently result in lower ranks across languages compared to the original prompts, indicating that entity representations are more aligned in the common conceptual space, resulting in more consistent crosslingual factual prediction.

results for LLaMA (8B) and OLMo (13B), with additional models included in Appendix §C.

As we observe in Figure 6, the solid lines representing SUBSUB and SUBINJ consistently yield lower ranks across layers compared to the dashed lines from the original prompts. This trend is especially noticeable in later layers, where factual decoding typically occurs. These observations indicate that both prompting strategies strengthen entity representation alignment across languages. The improved alignment of objects suggests that subject alignment – achieved through input prompts with substitutions/injections – engages reliable pivot-language processing in the shared conceptual space and then propagates to better object alignment, reducing inconsistencies that possibly arise from misaligned subjects.

Overall, these results confirm that SUBSUB and SUBINJ enhance crosslingual entity alignment at the *representation level*, therefore facilitating more accurate and consistent factual recall.

7 Conclusion

In this work, we report a set of converging observations that support the hypothesis that crosslingual factual consistency in multilingual LLMs is highly associated with entity-level alignment – the model's ability to map subject and object entities into a shared conceptual space across languages. In our experiments, we observe that when entity alignment fails, consistency rarely emerges, revealing alignment as a key factor for consistent multilingual factual recall. To address the inconsistency problem, we introduce two simple prompting strategies SUBSUB and SUBINJ that integrate English-translated subjects into queries, yielding substantial improvements in both factual recall accuracy and crosslingual consistency. Finally, our mechanistic analysis suggests that these prompt-based interventions help models better align entities within a shared conceptual space biased toward high-resource languages, e.g., English, thereby facilitating consistent factual recall.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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. arxiv preprint.

Leon Engländer, Hannah Sterz, Clifton A Poth, Jonas Pfeiffer, Ilia Kuznetsov, and Iryna Gurevych. 2024. M2QA: Multi-domain multilingual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6283–6305, Miami, Florida, USA. Association for Computational Linguistics.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16052–16106, Vienna, Austria. Association for Computational Linguistics.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. 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, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, and Others. 2025. Gemma 3 technical report.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, and Others. 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.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2025. MEXA: Multilingual Evaluation of English-Centric LLMs via Cross-Lingual Alignment. arxiv preprint: 2410.05873.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

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

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Ayyoob ImaniGooghari, Orgest Xhelili, Haotian Ye, Chunlan Ma, François Yvon, and Hinrich Schütze. 2025a. How transliterations improve crosslingual alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2417–2433, Abu Dhabi, UAE. Association for Computational Linguistics.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schütze. 2025b. Tracing multilingual factual knowledge acquisition in pretraining. arxiv preprint.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. arxiv preprint.

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. 2025. Lost in the mix: Evaluating LLM understanding of code-switched text. arXiv preprint: 2506.14012.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Nostalgebraist. 2020. Interpreting GPT: the logit lens. Blog post.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. COPEN: Probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5015–5035, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020.

LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do Multilingual LLMs Think In English? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 OLMo 2 Furious. arxiv preprint.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025. 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.

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.

Xiaolin Xing, Zhiwei He, Haoyu Xu, Xing Wang, Rui Wang, and Yu Hong. 2024. Evaluating knowledge-based cross-lingual inconsistency in large language models.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Relation	Number of Facts
applies_to_jurisdiction	79
capital	336
capital_of	212
continent	212
country_of_citizenship	60
developer	76
field_of_work	167
headquarters_location	51
instrument	46
language_of_work_or_name	108
languages_spoken	104
location_of_formation	66
manufacturer	35
native_language	130
occupation	46
official_language	602
owned_by	50
place_of_birth	35
place_of_death	79
religion	125
Total	**2,619**

Table 3: Number of facts per relation type.

A KLAR Dataset Details

The statistics of the KLAR dataset (Wang et al., 2025) are presented in Table 3. KLAR is based on BMLAMA17 (Qi et al., 2023) with some minor modifications to improve the applicability to autoregressive models. We use **2,619** facts grouped into **20** relation categories.

B Supplementary Results

In §5, we propose two simple but effective strategies – SUBSUB and SUBINJ – which inject or substitute the subject entity in prompts with its English translation. These interventions yield substantial improvements in both factual recall accuracy and consistency, particularly for smaller multilingual models and English-centric models. To better understand the role of English (the pivot language) in these gains, we conduct a complementary study using **Spanish** and **Japanese** translations of subject entities instead using four models: LLaMA (1B), LLaMA (8B), OLMo (1B), and OLMo (13B). Table 4 presents the aggregated performance; Figures 8, 9, 10 report the consistency across language pairs using English, Spanish, and Japanese subject translations, respectively.

Japanese underperforms, especially in English-centric models. Using Japanese as the pivot language results in limited or even negative gains. For example, in OLMo (1B), SUBSUB with Japanese degrades performance compared to the base prompt. These findings suggest that when the pivot language differs substantially from the model's conceptual embedding space, entity alignment becomes harder, leading to poorer factual recall accuracy and consistency. This highlights the critical role of conceptual space alignment in driving factual recall.

Conceptual space bias explains differences across pivot languages. These observations support our broader hypothesis: the model's ability to project entities into a shared, language-agnostic conceptual space facilitates consistent factual recall. In practice, however, this space is not neutral – it is biased toward high-resource or pretraining-dominant languages like English. As a result, using English-translated subjects helps the model activate this space more effectively than using other languages like Japanese. Spanish, which shares the script and extensive lexical overlap, therefore offers comparable improvements with English.

C Mechanistic Interpretability: Additional Results

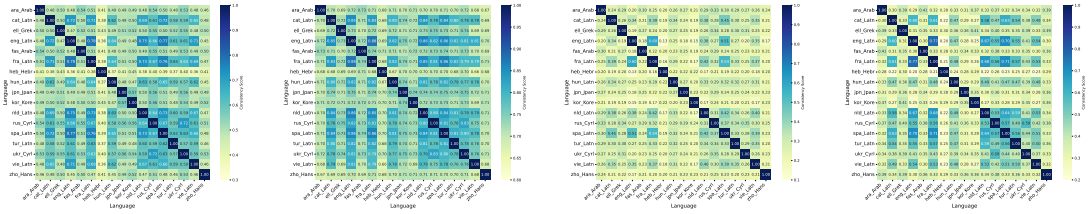
As mentioned in §6, we use Logit Lens analysis to examine how SUBSUB and SUBINJ influence internal model representations. Here, we present the results on LLaMA (1B) and OLMo (1B).

Figure 11 shows the rank of the target object token and its equivalents across six languages under original prompts, SUBSUB, and SUBINJ. We observe consistent patterns with the larger models: both interventions reduce object ranks across layers, this confirms that our findings generalize across different model scales.

In addition to rank analysis, we also examine probability curves for the target object token in each language under the three prompting conditions (original prompt, SUBSUB, and SUBINJ). As shown in Figure 12, the interventions consistently increase the predicted probability of the correct object across layers. This complements the rank analysis and further demonstrates that providing English subject cues enhances crosslingual entity alignment, making the correct object more accessible during decoding.

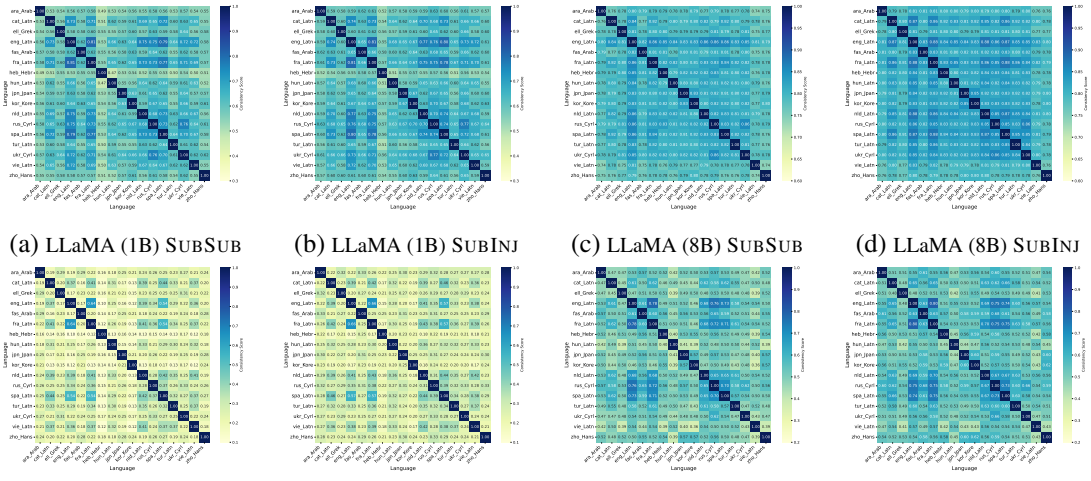
Model	Lang	Recall (ACC)					Consistency (CO)				
		Base	SUBSUB	SUBINJ	↑SUBSUB (%)	↑SUBINJ (%)	Base	SUBSUB	SUBINJ	↑SUBSUB (%)	↑SUBINJ (%)
LLaMA (1B)	eng	0.53	0.60	**0.61**	11.7	14.9	0.54	0.61	**0.63**	13.2	17.8
	jpn	0.53	0.41	**0.55**	-23.3	2.8	0.54	0.50	**0.57**	-7.2	5.2
	spa	0.53	0.56	**0.59**	4.3	10.6	0.54	0.59	**0.61**	9.2	14.4
LLaMA (8B)	eng	0.71	0.75	**0.76**	5.6	7.2	0.74	0.80	**0.82**	7.7	10.0
	jpn	0.71	0.66	**0.73**	-6.9	3.4	0.74	0.73	**0.78**	-1.9	5.0
	spa	0.71	0.72	**0.75**	1.5	5.7	0.74	0.77	**0.80**	4.3	8.2
OLMo (1B)	eng	0.27	0.28	**0.31**	2.4	11.2	0.27	0.24	**0.28**	-8.7	4.8
	jpn	**0.27**	0.15	0.27	-46.2	-2.7	**0.27**	0.23	0.25	-13.4	-5.1
	spa	0.27	0.25	**0.29**	-8.6	5.0	0.27	0.24	**0.27**	-7.8	1.7
OLMo (13B)	eng	0.43	0.54	**0.56**	25.9	31.6	0.39	0.52	**0.55**	35.3	43.9
	jpn	0.43	0.27	**0.45**	-36.2	4.7	0.39	0.41	**0.43**	5.8	10.5
	spa	0.43	0.49	**0.53**	15.1	24.1	0.39	0.50	**0.53**	30.2	36.9

Table 4: Performance comparison of the proposed remedies SUBSUB and SUBINJ for improving entity-level alignment, using English, Japanese, and Spanish translations of subjects. Both methods improve recall and consistency when the intervention leverages translations from **high-resource languages** such as English and Spanish, indicating that the alignment benefits from the conceptual space biased towards these languages.



(a) LLaMA (1B) (b) LLaMA (8B) (c) OLMo (1B) (d) OLMo (13B)

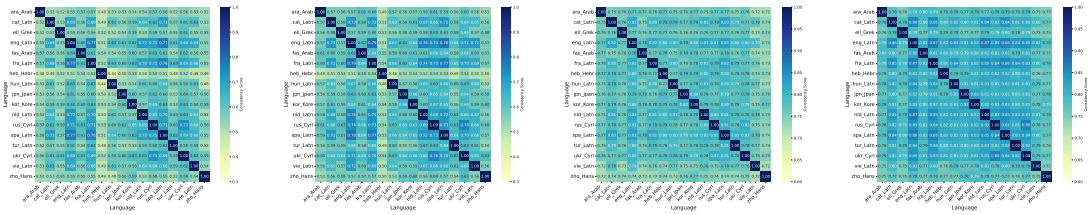
Figure 7: Factual recall consistency across language pairs under **baseline prompting**, with language-specific factual queries issued in their original languages.



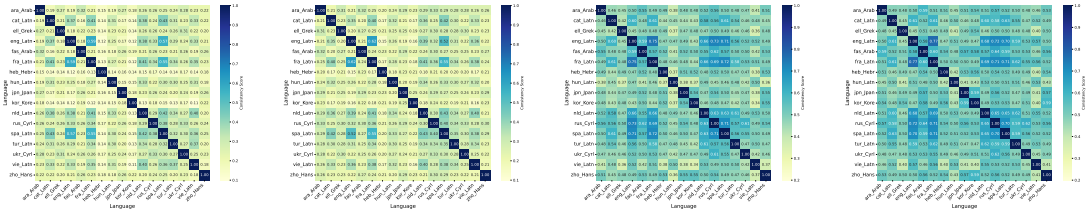
(a) LLaMA (1B) SUBSUB (b) LLaMA (1B) SUBINJ (c) LLaMA (8B) SUBSUB (d) LLaMA (8B) SUBINJ

(e) OLMo (1B) SUBSUB (f) OLMo (1B) SUBINJ (g) OLMo (13B) SUBSUB (h) OLMo (13B) SUBINJ

Figure 8: Factual recall consistency across language pairs using **English** translations of subjects in SUBSUB and SUBINJ.

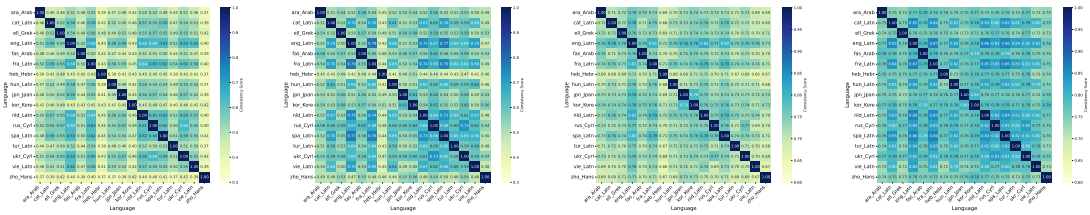


(a) LLaMA (1B) SUBSUB (b) LLaMA (1B) SUBINJ (c) LLaMA (8B) SUBSUB (d) LLaMA (8B) SUBINJ

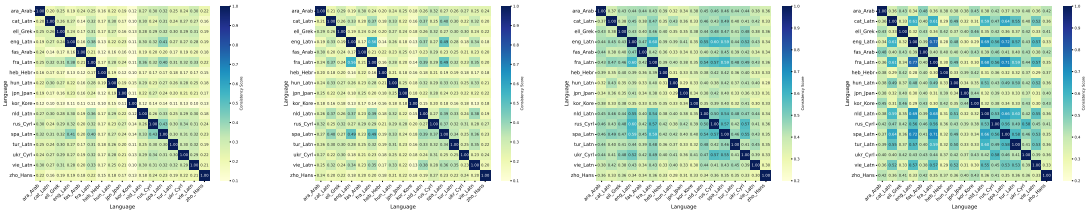


(e) OLMo (1B) SUBSUB (f) OLMo (1B) SUBINJ (g) OLMo (13B) SUBSUB (h) OLMo (13B) SUBINJ

Figure 9: Factual recall consistency across language pairs using **Spanish** translations of subjects in SUBSUB and SUBINJ.

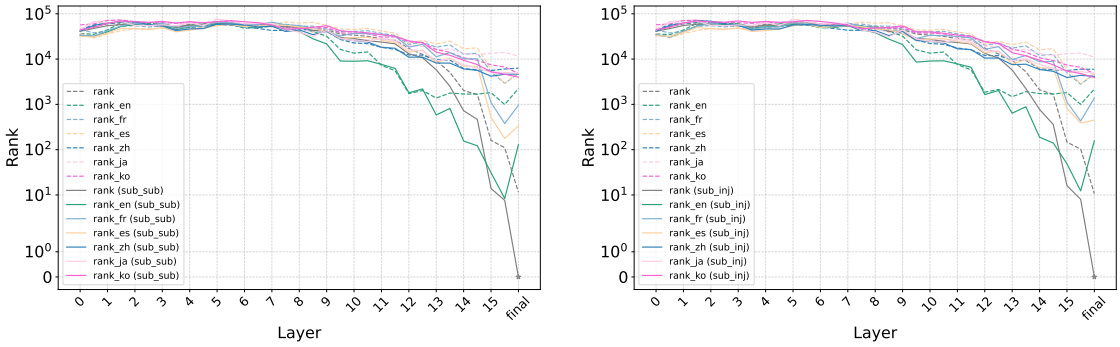


(a) LLaMA (1B) SUBSUB (b) LLaMA (1B) SUBINJ (c) LLaMA (8B) SUBSUB (d) LLaMA (8B) SUBINJ

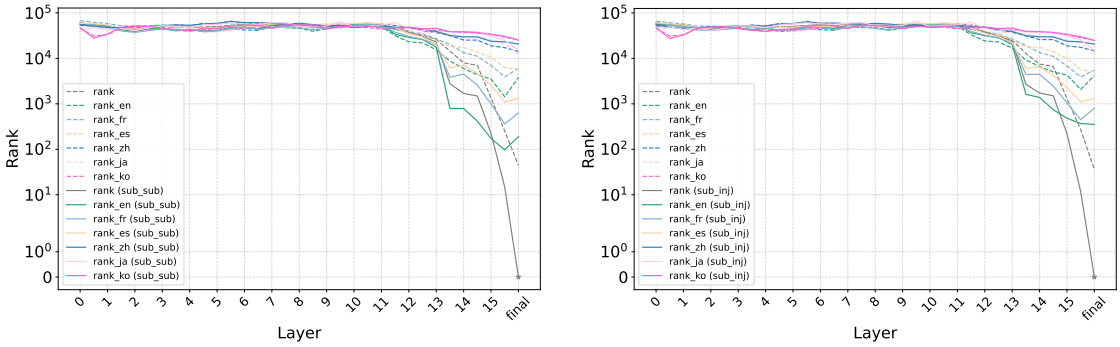


(e) OLMo (1B) SUBSUB (f) OLMo (1B) SUBINJ (g) OLMo (13B) SUBSUB (h) OLMo (13B) SUBINJ

Figure 10: Factual recall consistency across language pairs using **Japanese** translations of subjects in SUBSUB and SUBINJ.

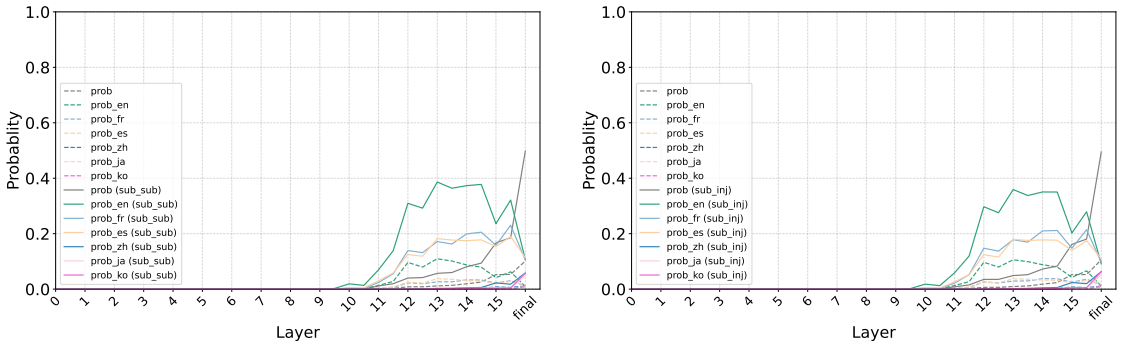


(a) **LLaMA (1B)**: Left – with SUBSUB; Right – with SUBINJ.

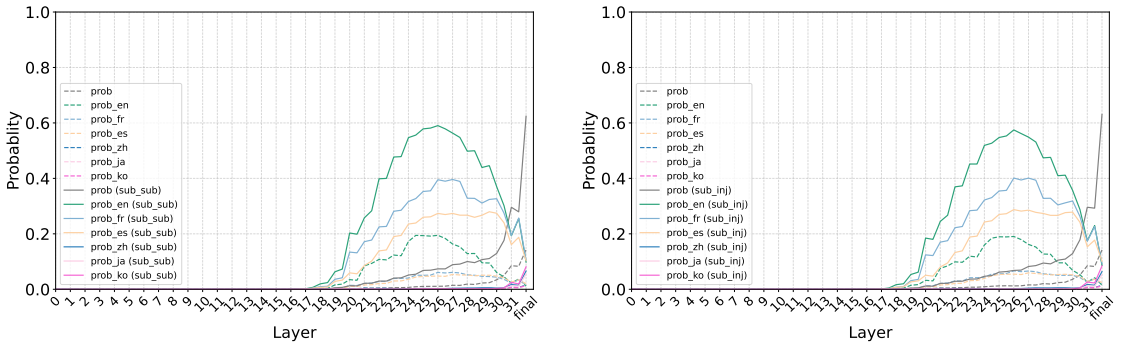


(b) **OLMo (1B)**: Left – with SUBSUB; Right – with SUBINJ.

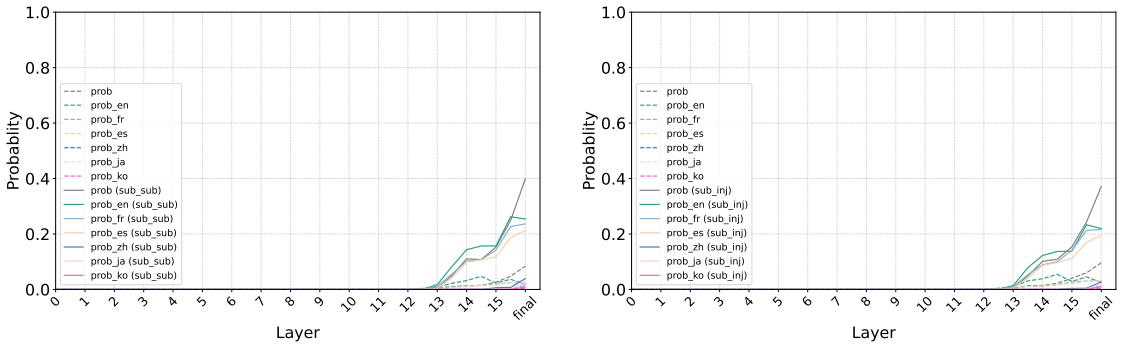
Figure 11: **Logit Lens analysis of object token ranks across model layers.** We plot the rank of the target object token (lower is better) in its input language, along with its equivalents in six other languages, using lines of different colors. Dashed lines represent the original prompts, while solid lines correspond to prompts modified with SUBSUB and SUBINJ. Both interventions consistently result in lower ranks across languages compared to the original prompts, indicating entity representations are more aligned in the common conceptual space, and consequently, more consistent crosslingual factual prediction.



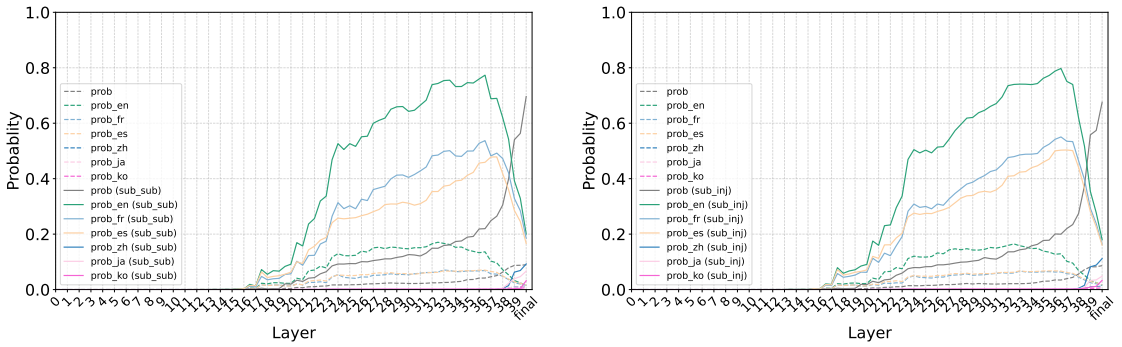
(a) **LLaMA (1B)**: Left – with SUBSUB; Right – with SUBINJ.



(b) **LLaMA (8B)**: Left – with SUBSUB; Right – with SUBINJ.



(c) **OLMo (1B)**: Left – with SUBSUB; Right – with SUBINJ.



(d) **OLMo (13B)**: Left – with SUBSUB; Right – with SUBINJ.

Figure 12: **Logit Lens analysis of object token probabilities across model layers.** We plot the probability of the target object token in its input language, along with its equivalents in six other languages, using lines of different colors. Dashed lines represent the original prompts, while solid lines correspond to prompts modified with SUBSUB and SUBINJ. Both interventions consistently result in higher probabilities across languages compared to the original prompts, indicating entity representations are more aligned in the conceptual space, and consequently, more consistent crosslingual factual prediction.