Tracing Multilingual Factual Knowledge Acquisition in Pretraining

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are capable of recalling multilingual factual knowledge present in their pretraining data. However, most studies evaluate only the final model, leaving the development of *factual recall* and *crosslingual consistency* throughout pretraining largely unexplored. In this work, we trace how factual recall and crosslingual consistency evolve during pretraining, focusing on OLMo-7B as a case study. We find that both accuracy and consistency improve over time for most languages. We show that this improvement is primarily driven by the *fact frequency* in the pretraining corpus: more frequent facts are more likely to be recalled correctly, regardless of language. Yet, some low-frequency facts in non-English languages can still be correctly recalled. Our analysis reveals that these instances largely benefit from crosslingual transfer of their English counterparts – an effect that emerges predominantly in the early stages of pretraining. We pinpoint two distinct pathways through which multilingual factual knowledge acquisition occurs: (1) ***frequency-driven learning***, which is dominant and language-agnostic, and (2) ***crosslingual transfer***, which is limited in scale and typically constrained to relation types involving named entities. We will release our code to facilitate further research.

1 Introduction

Despite being predominantly trained on English-centric data, LLMs exhibit surprisingly strong multilingual capabilities across a wide range of tasks (Jiang et al., 2023; Touvron et al., 2023; Zhang et al., 2024; Zhao et al., 2025). Notably, they can recall factual knowledge in multiple languages (Petroni et al., 2019; Jiang et al., 2020; Kassner et al., 2021). However, these models frequently exhibit *crosslingual inconsistencies* – answering a factual query correctly in one language but failing to do so in another (Qi et al., 2023; Chua et al., 2025; Wang et al., 2025). Although bilinguals typically recall information more effectively when the language of encoding matches the language of retrieval, they can usually recall factual knowledge learned in one language using their other proficient language (Marian and Neisser, 2000; Chung et al., 2019) – highlighting a flexibility that contrasts with the inefficiencies seen in LLMs. Understanding this discrepancy requires deeper insight into how multilingual factual knowledge is acquired.

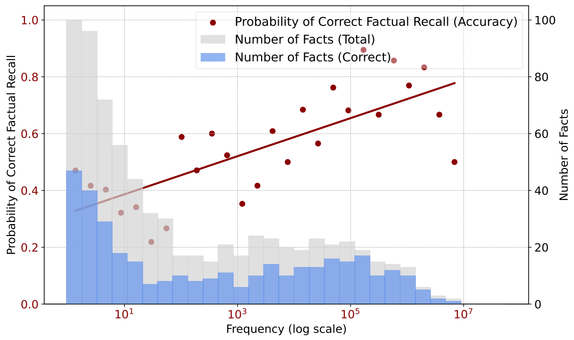

Figure 1: Relationship between fact frequency and factual recall in **Catalan**. High-frequency facts are more likely to be correctly recalled, indicating the effect of *frequency-based learning*. Meanwhile, the correct recall of some low-frequency facts suggests the influence of *crosslingual transfer* from other languages.

While prior work has investigated mechanisms of (multilingual) factual recall (Geva et al., 2023; Zhao et al., 2024; Fierro et al., 2024; Liu et al., 2025) and analyzed sources of crosslingual inconsistency (Qi et al., 2023; Wang et al., 2025), these studies have largely focused on *final models*, drawing conclusions solely from the end of pretraining. As a consequence, the developmental process by which LLMs acquire factual knowledge across languages remains poorly understood.

To address this gap, we trace the dynamics of multilingual factual recall and crosslingual consistency throughout pretraining. Rather than treating factual recall as a static outcome, we analyze its emergence across checkpoints using OLMo-

7B (Groeneveld et al., 2024), an English-centric decoder-only LLM pretrained on Dolma (Soldaini et al., 2024). Our analysis evaluates both accuracy within individual languages and consistency across languages for facts that are parallel in all languages.

In addition, we investigate the key factors that contribute to correct multilingual factual recall. Prior work has shown that the frequency of an instance can significantly influence performance relating to it, including factual prediction (Razeghi et al., 2022; Elazar et al., 2023; McCoy et al., 2024; Merullo et al., 2025). Motivated by these findings, we hypothesize that *fact frequency* in the pretraining corpus plays a central role in multilingual factual recall. To test this, we compute the frequency of each fact and systematically link it to factual recall across languages and pretraining stages.

We summarize the key findings of this paper:

(i) **The capacity for multilingual factual recall develops progressively during pretraining** (§4). English and languages distant from English converge in early stages, while languages more similar to English (e.g., those sharing the Latin script) continue to improve with extended pretraining.

(ii) **The correctness of factual recall is largely explained by a single factor: fact frequency in the pretraining corpus** (§5). High-frequency facts are consistently recalled more accurately across languages (e.g., Catalan in Figure 1). In addition, this frequency-correctness relationship emerges early and strengthens throughout pretraining.

(iii) **Some low-frequency facts in non-English languages are recalled correctly mainly via crosslingual transfer** (§6). High-frequency counterparts in English mainly enable these cases. However, the scale of transfer is limited and constrained to certain relation types.

2 Related Work

Multilingual Factual Recall and Consistency Several studies have investigated the factual knowledge stored in models through knowledge probing. Jiang et al. (2020) and Kassner et al. (2021) assess factual recall by translating English prompts into multiple languages, revealing notable performance disparities across languages. Yin et al. (2022) extend this analysis to region-specific commonsense knowledge, finding that the best-performing language for querying facts about a country (e.g., China) is often English rather than its native language (e.g., Chinese), indicating the English-centric bias of models. Building on multilingual probing studies, Qi et al. (2023) and Aggarwal et al. (2025) investigate crosslingual consistency and find that LLMs often return different answers for equivalent queries in different languages. Wang et al. (2025) further explore the underlying causes of these inconsistencies through mechanistic interpretability, revealing how internal representations contribute to divergent outputs across languages. Following this line of research, our work traces the development of factual recall and crosslingual consistency throughout pretraining, shedding light on how these capabilities emerge and evolve.

Pretraining Trajectory Investigation Several studies have investigated how Transformer-based models (Vaswani et al., 2017) acquire linguistic or task-specific knowledge during different phases of pretraining, in both monolingual (Choshen et al., 2022; Xia et al., 2023; Müller-Eberstein et al., 2023; Chen et al., 2024) and multilingual settings (Blevins et al., 2022; Wang et al., 2024). A concurrent study by Merullo et al. (2025) most closely resembles our work; they demonstrate that fact frequency is a strong predictor of both factual recall and the emergence of linear factual representations (e.g., subject-to-object mappings via linear transformation) (Hernandez et al., 2024). However, their analysis is conducted in a purely monolingual context. In contrast, our work examines multilingual factual knowledge acquisition and shows that while fact frequency remains a key driver of factual recall, crosslingual knowledge transfer provides additional – albeit limited – benefits in enhancing multilingual factual recall.

3 Experiment Setups

3.1 Languages and Model Checkpoints

Languages We consider 12 languages that span 6 language families and use 7 different scripts: Arabic (**ara_Arab**), Catalan (**cat_Latn**), Chinese (**zho_Hans**), English (**eng_Latn**), French (**fra_Latn**), Greek (**ell_Grek**), Japanese (**jpn_Jpan**), Korean (**kor_Kore**), Russian (**rus_Cyrl**), Spanish (**spa_Latn**), Turkish (**tur_Latn**), Ukrainian (**ukr_Cyrl**).[1]

[1] Some languages, e.g., Ukrainian, are much less resourced than others, according to our exploration of the multilingual

Model Checkpoints We use the open-source OLMo-1.7 7B model (Groeneveld et al., 2024) (referred to as OLMo) in our study. OLMo is a decoder-only LLM pretrained on Dolma (Soldaini et al., 2024), an English-centric corpus with some multilingual coverage. To capture the dynamics of factual knowledge acquisition throughout pretraining, we select model checkpoints at two granularities. Based on preliminary experiments showing that changes are more pronounced in the early pretraining stages, we include checkpoints every 1,000 steps from step 0 to step 50,000. Beyond 50,000 steps, we consider every 5,000 steps up to step 400,000. This setup enables us to trace the model's development from initialization to a mature stage with good multilingual capability (trained on approximately 1.7T tokens).

3.2 Multilingual Factual Dataset

We use KLAR (Wang et al., 2025), a multilingual factual knowledge probing dataset, for our investigation. We use **1,197** facts grouped into **12** relation categories (cf. Table 2 in §A). Each fact is represented as a triple of subject, relation, and object. KLAR also provides a prompt template for each relation in each language, structured as "`<Question> The answer is:`". For example, for triple (*France, capital, Paris*), the template will then be expanded as "*Where is France's capital located? The answer is:*", with expected answer "*Paris*" in English. All facts and prompt templates are available in all 12 languages. We therefore transform each fact into a query q_i^l with expected answer o_i^l in language l; for each fact i, q_i^l and $q_i^{l'}$ are translations of the same query in languages l and l'. We denote the resulting set of queries as Q.

3.3 Evaluation

To evaluate **consistency**, we compute the overlapping ratio of correct predictions, following Jiang et al. (2020) and Wang et al. (2025). Since OLMo is an English-centric model due to the predominance of English in Dolma's documents (cf. §J), we treat English as a *reference language* and compute how consistent the predictions from other languages are compared to predictions made in English:[2]

$$\mathrm{CO}(l) = \frac{\sum_i^{|Q|} \mathbf{1}(\mathcal{M}(q_i^l) = o_i^l \wedge \mathcal{M}(q_i^{\mathrm{eng}}) = o_i^{\mathrm{eng}})}{\sum_i^{|Q|} \mathbf{1}(\mathcal{M}(q_i^l) = o_i^l \vee \mathcal{M}(q_i^{\mathrm{eng}}) = o_i^{\mathrm{eng}})}$$

where q_i^{eng} and o_i^{eng} are the query and expected answer for the ith query in English, $\mathbf{1}(\cdot)$ is the indicator function, and $\mathcal{M}(\cdot)$ is the LLM's prediction function. When assessing correctness ($\mathcal{M}(q_i^l) = o_i^l$), we rely on the model's *complete generation*, checking whether it contains o_i^l. We depart here from previous work (Geva et al., 2023; Qi et al., 2023; Hernandez et al., 2024) that just checks the first predicted token, which can be misleading due to ambiguity and tokenization issues.[3] We also compute the per language **accuracy**: $\mathrm{ACC}(l) = \frac{\sum_i^{|Q|} \mathbf{1}(\mathcal{M}(q_i^l) = o_i^l)}{|Q|}$ which allows us to trace how well factual recall is performed.

3.4 Fact Frequencies

We approximate a fact's frequency by counting the number of documents where **its subject and object co-occur** in the pretraining corpus. This co-occurrence-based approximation has been widely used and shown to be reliable (Elsahar et al., 2018; Elazar et al., 2023; Merullo et al., 2025; Liu et al., 2025). For some languages, this approximation is fairly accurate due to the uniqueness of their scripts – for example, the subject-object pair (法国, 巴黎) in Chinese is unlikely to appear in texts from other languages. However, ambiguity arises in languages that share scripts, such as English and French. The same pair (*France, Paris*), for instance, may appear in either language, resulting in an *aggregated frequency* count shared across both. We analyze the impact of this identical-fact effect and show that it does not compromise the robustness of our findings (cf. §I). To efficiently obtain these co-occurrence counts, we use the ElasticSearch API provided by **WIMBD** (Elazar et al., 2024), a tool designed for scalable search and frequency analysis over large corpora.[4] All fact frequencies in our analysis are computed over the Dolma v1.7 corpus (Soldaini et al., 2024) used to pretrain OLMo, by measuring the number of subject-object co-occurrences for each fact in KLAR.

4 Multilingual Factual Recall Dynamics

We begin our analysis by tracing how factual recall performance evolves throughout pretraining

coverage of Dolma (Soldaini et al., 2024) (cf. §J).

[2] We present a complementary investigation of holistic crosslingual consistency across all language pairs in §C.

[3] Even though the first token is correct, the final prediction can be wrong because the object is split into multiple tokens. For example, "Antwerp" and "Antananarivo" share the same first token "Ant". It is therefore ambiguous which city the model is trying to generate based on just the token "Ant".

[4] A public demo of WIMBD is available at: `https://wimbd.apps.allenai.org/`.

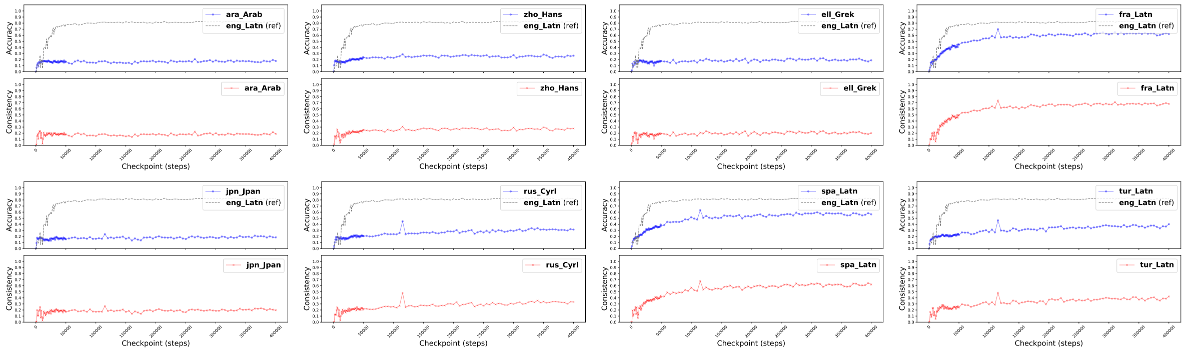

Figure 2: Factual accuracy (ACC) and crosslingual consistency (CO). While factual knowledge is rapidly acquired during the early stages of pretraining and is reasonably high in many languages, a substantial performance gap remains between English and most other languages, highlighting the limitations of crosslingual knowledge transfer.

across different languages. Specifically, we examine both *accuracy* and *crosslingual consistency* at each checkpoint of OLMo (cf. §3.1) using the KLAR dataset. Figure 2 summarizes these results for eight languages (see §B for full results).

Crosslingual consistency is tightly coupled with non-English performance. We observe that the trajectory of crosslingual consistency in each language $l \neq$ eng_Latn closely mirrors its own factual accuracy throughout pretraining. This suggests that consistency is primarily driven by whether the fact is correctly recalled in l, which almost always implies that it is also recalled in English. The implication is twofold. (1) For non-English languages, the consistency of a language (CO) is effectively gated by its performance (ACC). (2) The limited capability of the model to transfer knowledge from English to other languages, referred to as the *crosslingual knowledge barrier* (Chua et al., 2025), is a persistent problem throughout pretraining.

Factual knowledge is acquired rapidly in early pretraining phases. We observe that factual recall performance (ACC) improves very quickly in the early stages of pretraining for many languages. For example, English reaches approximately 80% accuracy after only 50K steps (roughly 209B tokens), with minimal gains beyond that point. This indicates that factual knowledge is acquired rapidly early and does not substantially benefit from further pretraining steps. While longer pretraining is known to improve other capabilities of LLMs (Kaplan et al., 2020; Le Scao et al., 2022; Xiong et al., 2024), factual recall appears to rely on simpler mechanisms gained in early-stage training, likely tied to memorization of frequent co-occurrences, for which we give empirical evidence in §5.

Script plays a more important role than language family in sustained improvements. Languages such as ara_Arab, jpn_Jpan, and kor_Kore, which neither use the Latin script nor belong to the Indo-European family, reach early saturation in performance – typically even before 2K steps. In contrast, Latin-script languages such as cat_Latn, fra_Latn, and spa_Latn, continue to improve with more training steps. Interestingly, ell_Grek, despite being an Indo-European language, saturates early as well, whereas tur_Latn, from the Turkic family, benefits from extended pretraining. This pattern suggests that surface features like script similarity are more influential for possible crosslingual knowledge transfer than deeper typological relationships, as we further investigate in §6.

5 Fact Frequency As Predictor

A notable observation in §4 is that factual recall performance (ACC) rapidly converges for many languages, including English. This suggests that the model acquires much of its factual knowledge in the early stages of pretraining and is able to recall it reliably when appropriately prompted (cf. §3.2). We hypothesize that this behavior reflects a form of memorization, where frequent exposure to specific facts in the pretraining corpus enables the model to retrieve them accurately. To investigate this, we approximate the frequency of all facts in the KLAR dataset (cf. §3.4) and analyze the relationship between frequency and factual recall performance both "**globally**" – across all languages – and "**locally**" – within individual languages.

5.1 Global Results Across All Languages

We analyze the relationship between fact frequency (in log scale) and probability of correct factual recall across six OLMo checkpoints: 5K, 10K, 30K,

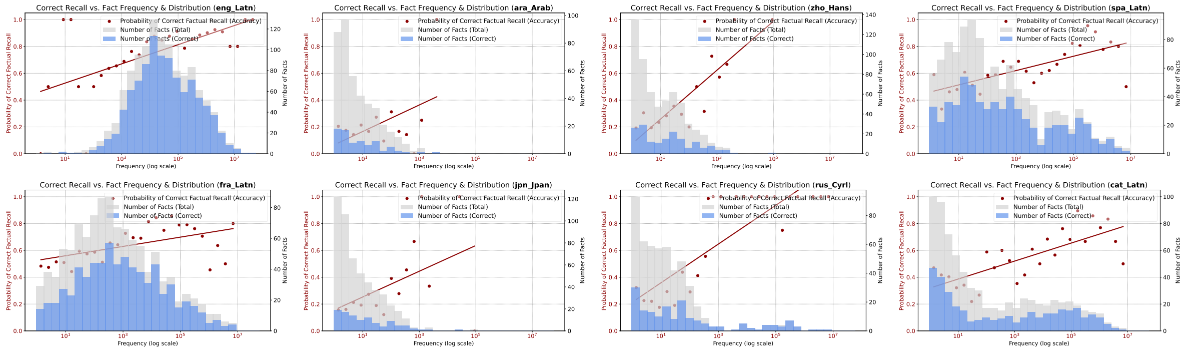

Figure 3: Relationship between fact frequency and the probability of correct factual recall. A consistent upward trend across individual languages indicates that higher-frequency facts are more likely to be recalled by the model.

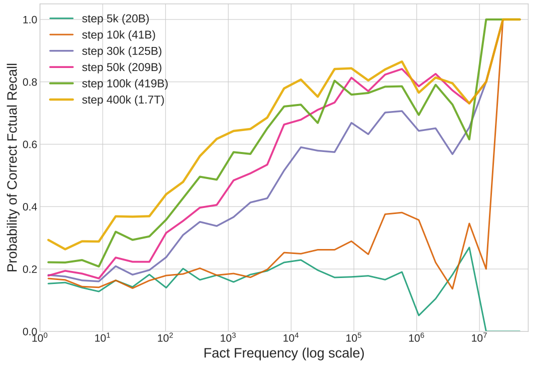

Figure 4: Relationship between fact frequency and factual recall for all languages and six pretraining checkpoints. High-frequency facts are more likely to be correctly recalled than rare ones. This frequency-correctness correlation emerges early in pretraining and becomes more pronounced over time.

50K, 100K and 400K. The results are displayed in Figure 4. Results for more checkpoints are reported in §D.1.

Fact frequency strongly predicts factual recall performance. At the 400K-step checkpoint (corresponding to approximately 1.7T tokens), we observe a strong positive correlation between the fact log frequency and the probability of correct factual recall, with a Pearson correlation coefficient of $r = 0.93$ ($p < 0.001$). This indicates a robust linear relationship between the two variables and supports our hypothesis that fact frequency in the pretraining corpus is a key determinant of factual recall performance across languages.

This correlation emerges early in pretraining. While the 5K-step and 10k-step checkpoints (around 20B and 41B tokens, respectively) show weak correlation, the 30K-step checkpoint (around 125B tokens) has Pearson coefficients $r = 0.95$, indicating strong correlation. Together with the high factual recall accuracy observed in early checkpoints (cf. Figure 2), these results suggest that the model is exposed to and memorizes many high-frequency facts early in pretraining, enabling accurate recall even before large-scale exposure, aligned with findings from Merullo et al. (2025).

5.2 Analysis per Language

We further investigate whether the relationship between fact frequency and factual recall accuracy holds consistently across individual languages. We focus on the 400k-step checkpoint.

High-frequency facts are more likely to be correctly recalled within individual languages. Figure 3 shows the distribution of fact frequencies and corresponding factual recall probabilities for 8 representative languages (results for additional languages are in §D.2). Across all cases, we observe a clear trend: facts that occur more frequently in the pretraining corpus are more likely to be correctly recalled. This pattern is not limited to English; languages such as rus_Cyrl exhibit particularly strong effects – for instance, when fact frequency exceeds 10^3, the model recalls the fact with near-perfect accuracy. Similar trends are observed in other languages as well, suggesting that fact frequency plays a consistently central role in determining factual recall performance across languages.

5.3 Recall Prediction with Frequencies

We observed in §5.2 that the relationship between fact frequency and factual recall holds consistently across individual languages. This naturally leads to a further question: **Can the recallability of a fact be reliably predicted solely based on its frequency within a given language?** To answer this, we construct a simple frequency-based classifier for each language and evaluate its effectiveness. Again, we focus on the 400k-step checkpoint.

Formally, for each language l, we define a dataset $\mathcal{D}_l = \{(f_i^l, y_i^l)\}_{i=1}^N$, where $f_i^l \in \mathbb{Z}_{\geq 0}$

Lang	Threshold	Accuracy	FN
ara_Arab	3485	0.83	209
cat_Latn	2506	0.63	384
ell_Grek	483	0.84	190
eng_Latn	108	0.82	7
fra_Latn	19	0.64	134
jpn_Jpan	352	0.82	212
kor_Kore	402	0.80	238
rus_Cyrl	370	0.72	330
spa_Latn	12	0.60	169
tur_Latn	3068	0.64	373
ukr_Cyrl	385	0.79	248
zho_Hans	502	0.75	296

Table 1: Best threshold, accuracy, and false negatives when using fact frequency as a predictor of factual recall. We interpret FN as **s**urprising **c**orrect **l**ow-**f**requency **p**redictions (***SCLFP***) – predictions that are correct even though the underlying fact frequency is low. Good accuracy on assessing fact frequency as a predictor for correct fact recall is achieved for most languages with this classifier as shown in column "Accuracy".

is the frequency of fact i, and $y_i^l \in \{0, 1\}$ indicates whether the model correctly recalled the fact (1 if correct, 0 otherwise). $\mathbb{Z}_{\geq 0}$ is the set of positive integers including 0. We then define a threshold-based classifier $h_t^l(f)$ for each language as: $h_t^l(f) = \begin{cases} 1, & \text{if } f \geq t \\ 0, & \text{otherwise} \end{cases}$. The optimal threshold t_l^* in each language is selected to maximize classification accuracy:

$$t_l^* = \arg\max_{t \in \mathbb{Z}_{\geq 0}} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(h_t^l(f_i^l) = y_i^l\right)$$

where $\mathbf{1}(\cdot)$ is the indicator function. To better understand the classification behavior, we also compute the number of false negatives (FN) under the optimal threshold, as these facts are also correctly predicted but with low frequencies.[5] Table 1 presents the classification performance.

Fact frequency serves as a strong predictor of factual recall for many languages. Across all languages, the threshold-based classifier achieves accuracy above 0.6, indicating performance much better than random guessing. A closer inspection reveals that all languages with relatively lower accuracy, i.e., fra_Latn, spa_Latn, tur_Latn, and cat_Latn, use the Latin script, with no exceptions. In contrast, languages using non-Latin scripts consistently achieve higher accuracy.[6] We hypothesize that this pattern stems from extensive crosslingual transfer from English to other Latin-script languages. As a result, many low- or mid-frequency facts in these languages may still be correctly recalled, likely due to shared vocabulary and lexical overlap, as also shown by Qi et al. (2023). This transfer effect tends to shift the optimal classification threshold downward, enabling the threshold-based classifier to correctly predict low-frequency facts more often than expected.

All languages but English exhibit large false negative rates. This is particularly clear in languages using non-Latin scripts, such as ara_Arab and ukr_Cyrl, where the classifier fails to capture many low-frequency facts that are in fact recalled correctly by the model. Even in Latin-script languages – where the accuracy is relatively lower than in other languages due to the reasons noted above – we still observe a substantial number of false negatives. English stands out as the only language with few false negatives, because of the generally high fact frequencies. This consistent trend across languages suggests that many low-frequency facts are correctly recalled, motivating a closer examination of such cases. We further investigate them in §6.

6 Investigation of Transfer Effect

We observed a substantial number of false negatives when using frequency as a predictor in §5.3, particularly for languages that use non-Latin scripts. This is counterintuitive given the strong role frequency typically plays in factual recall. We hypothesize that these cases are due to the **crosslingual transfer** effect – factual knowledge is primarily learned in English and is successfully transferred to other languages. In the following sections, we present a detailed analysis of these false negatives identified in §5.3 – which we will refer to as **s**urprisingly **c**orrect **l**ow-**f**requency **p**redictions (***SCLFP***s).

6.1 Relation Type Distribution

We hypothesize that facts that involve named entities or shared vocabulary are easier to transfer across languages – e.g., the subject-object pair France-Paris is easy to transfer from English to French since the two named entities are identical

[5] Other error types are not the primary focus of our further analysis presented in the main content. For example, the cause of false positives may be due to (1) insufficient exposure to the fact despite its high frequency, or (2) sensitivity to the specific prompt used for evaluation. We present an analysis of the classifier in §E and complete error breakdown in §I.

[6] We conduct a sensitivity analysis on the classifier in §E and show it is more robust in non-Latin-script languages.

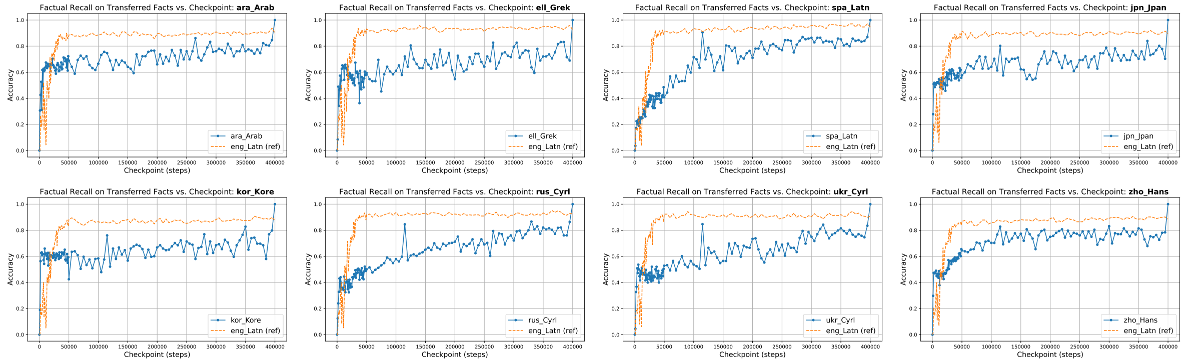

Figure 5: Dynamics of learning for ***SCLFP***s (surprisingly correct low frequency predictions, i.e., FNs in Table 1) across 8 languages. Crosslingual transfer emerges early in pretraining and continues to strengthen over time.

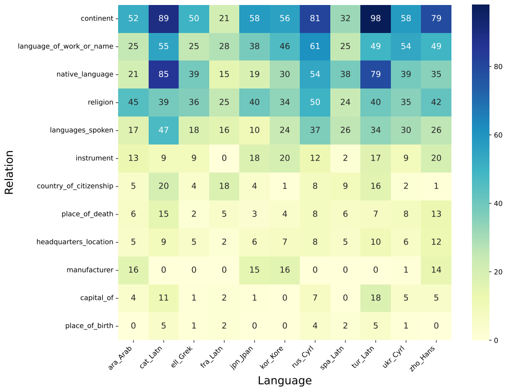

Figure 6: Distribution of ***SCLFP***s (surprisingly correct low frequency predictions) across relation types for each language. High ***SCLFP*** values are concentrated on relation types that involve only a small set of candidates – which are generally named entities.

in English and French. This intuition is grounded in how humans often rely on lexical similarity and recognizable entities when transferring knowledge. To investigate this, we group ***SCLFP***s in each language by their relation type, as shown in Figure 6.

***SCLFP*s are concentrated in relation types involving named entities.** This trend is especially pronounced in relations with a limited set of possible candidates, such as `continent` and `religion`. Languages that use a non-Latin script also benefit from named entity transfer, e.g., in `instrument` and `manufacturer` relations. This observation aligns with prior work showing that named entities are more easily transferred across script boundaries, particularly in encoder-only models (Imani et al., 2023; Liu et al., 2024a).

Latin-script languages benefit more broadly from crosslingual transfer. Compared to languages using other scripts, languages written in Latin script receive transfer benefits across a wider range of relations, such as `country_of_citizenship`. This is expected, as many Latin-script languages have substantial vocabulary overlap, leading to greater token-level similarity. Such overlap enables the transfer of identical or lexically similar entities – e.g., “Bulgària” in cat_Latn and “Bulgaristan” in tur_Latn. Moreover, higher token-level similarity in the context during pretraining can also facilitate the alignment, enhancing entity transfer (cf. §6.3).

6.2 Learning Progression

As shown in §4, the model acquires a substantial amount of factual knowledge during the early stages of pretraining. This raises a natural question: **Is crosslingual knowledge transfer similarly concentrated in the early stages, or does it continue throughout pretraining?** To explore this, we examine the learning trajectories of ***SCLFP***s across languages. Figure 5 illustrates how recall factual accuracy for ***SCLFP***s evolves over pretraing checkpoints for 8 languages (see full results in §G).

Extensive crosslingual transfer occurs during early pretraining. Across all languages, factual recall accuracy for ***SCLFP***s rapidly improves during the initial stages of pretraining. This trend is especially pronounced in languages that use non-Latin scripts. For example, ara_Arab, ell_Grek, and Kor_Kore reach over 60% accuracy within the first 20K steps, after which their performance plateaus or grows slowly, similar to the trend observed for in §4. These findings suggest that crosslingual transfer is not merely an emergent property of the final model, but rather a phenomenon that develops early in pretraining.

Many languages continue to benefit from transfer throughout pretraining. This is especially the case for languages using the Latin script, such as spa_Latn, which display a more gradual and continuous improvement. As discussed in §6.1, these languages benefit from crosslingual transfer across a broader range of relations, facilitated by extensive lexical overlap with other Latin-script languages. This broader scope of transferable content contributes to the prolonged learning curve. We also observe that rus_Cyrl and zho_Hans benefit from continued improvements over time, which could be attributed to the comparatively larger representation of Russian and Chinese texts in the pretraining corpus (cf. §J). Notably, ukr_Cyrl exhibits a learning curve that rapidly and closely aligns with rus_Cyrl, suggesting that transfer also occurs between other script-sharing languages (we show their consistency continues to improve in §C).

6.3 Similarity Dynamics

To better understand why certain languages, particularly those that do not use the Latin script, benefit from knowledge acquired in English, we analyze the evolution of cosine similarity between sentence-level representations of prompts (cf. §3.2) or fact pairs corresponding to ***SCLFP***s during pretraining. Specifically, we create fact pairs of ***SCLFP***s for each language, where every pair contains one prompt in that language and its counterpart in English. We then track the cosine similarity between these paired representations across checkpoints.[7] As a baseline, we also compute cosine similarities for ***UWLFP***s – **u**nsurprisingly **w**rong **l**ow-**f**requency **p**redictions identified in our frequency-based classification (cf. §5.2) – as well as for **all fact** pairs in each language. Figure 7 illustrates the progression of similarity scores over time for 6 languages (full results are available in §F).[8]

Similarity remains higher for *SCLFP*s than for *UWLFP*s. Across all languages, we observe a consistent trend: the cosine similarity for ***SCLFP***s quickly surpasses that of ***UWLFP***s. While both begin at comparable levels, a clear and sustained separation emerges after approximately 50K pretraining steps. This divergence suggests that the model aligns the representations of ***SCLFP***s with their English counterparts better than for ***UWLFP***s – facts that are similarly low-frequency but incorrectly predicted. These findings offer direct evidence of crosslingual knowledge transfer on ***SCLFP***s, benefiting from better alignment with English, spanning both language and script boundaries.

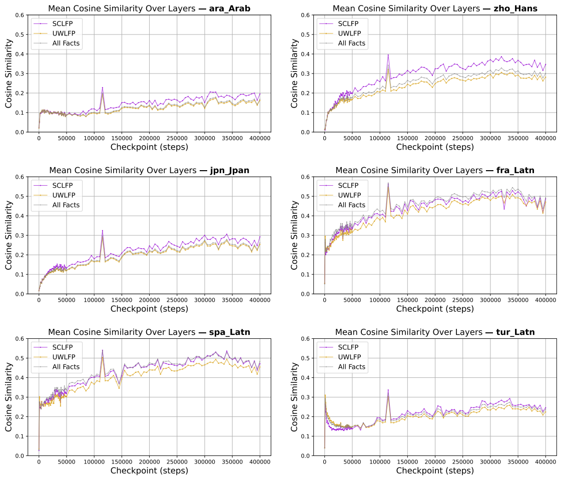

Figure 7: Mean cosine similarity between sentence-level representations of ***SCLFP***, ***UWLFP***, and all facts for each language paired with English during pretraining. All 6 languages exhibit consistently higher similarity for ***SCLFP*** than for ***UWLFP***, highlighting the emergence of crosslingual transfer through representation alignment.

Better alignment enables crosslingual transfer but does not guarantee correct recall. The consistently high similarity in Latin-script languages aligns with prior work showing that Transformer models tend to cluster representations based on shared script (Wen-Yi and Mimno, 2023; Liu et al., 2024b). However, improved alignment alone is not sufficient: for ***UWLFP***s, the model continues to better align them in pretraining, yet this does not lead to gains in recall accuracy (i.e., ***UWLFP***s are not learned). This suggests that beyond alignment, other factors – such as language-specific understanding/generation and instruction following abilities – also play a critical role in factual recall.

7 Conclusion

We investigate how multilingual factual recall and crosslingual consistency emerge during pretraining, using OLMo-7B as a case study. Our analysis shows that factual recall improves early and is primarily driven by fact frequency, regardless of language. However, some low-frequency facts in non-English languages can still be recalled, mainly due to crosslingual transfer from English – especially for relations that involve named entities. We therefore conclude that multilingual factual knowledge is gained through both frequency-driven learning and crosslingual transfer starting from early stages.

[7] We use the contextualized embedding of the final token as the sentence-level representation. Representations are extracted at each layer, and we report the mean cosine similarity computed by averaging similarities across all layers.

[8] To avoid inflated similarity, for each language, we filter out fact pairs where the object strings in that language and English are identical. Table 4 in §H provides statistics of fact pairs containing identical objects across languages.

Limitations

While this work contributes to emerging efforts in exploring multilingual knowledge acquisition during the pretraining process and contributes to understanding the mechanisms of acquisition, several limitations should be acknowledged.

First, our study focuses on the checkpoints of a single English-centric model as a case study. This choice is primarily due to the scarcity of open-source models that provide both intermediate checkpoints and detailed documentation of their pretraining corpora. We therefore echo Soldaini et al. (2024) and encourage greater transparency in the community, including the release of intermediate checkpoints and associated data. This would facilitate further research into knowledge acquisition dynamics and help deepen our understanding of LLM pretraining processes.

Second, our approximation of fact frequency in certain script-sharing languages may lack full accuracy. As discussed in §3.4 and §I, this is due to the difficulty in disambiguating language identity in shared-script corpora. While our findings suggest this issue does not significantly affect the overall results, future work could improve precision by applying language identification techniques, especially where computational resources permit.

Finally, although we analyze the dynamics of multilingual knowledge acquisition and identify two primary mechanisms – frequency-based learning and crosslingual transfer – we do not investigate the conditions under which each mechanism is most effective. Studying these underlying factors requires controlled manipulation of the pretraining corpus to observe causal effects, which falls beyond the scope of this work. Nonetheless, we regard this as a promising direction for future research.

References

Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang. 2025. Language models' factuality depends on the language of inquiry. *Preprint*, arXiv:2502.17955.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2025. Crosslingual capabilities and knowledge barriers in multilingual large language models. *Preprint*, arXiv:2406.16135.

Sheila Cira Chung, Xi Chen, and Esther Geva. 2019. Deconstructing and reconstructing cross-language transfer in bilingual reading development: An interactive framework. *Journal of Neurolinguistics*, 50:149–161. Cross-linguistic perspectives on second language reading.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2023. Measuring causal effects of data statistics on language model's 'factual' predictions. *Preprint*, arXiv:2207.14251.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2024. How do multilingual models remember? investigating multilingual factual recall mechanisms. *Preprint*, arXiv:2410.14387.

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.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

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.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

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

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

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.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. What language model to train if you have one million GPU hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yihong Liu, Runsheng Chen, Lea Hirlimann, Ahmad Dawar Hakimi, Mingyang Wang, Amir Hossein Kargaran, Sascha Rothe, François Yvon, and Hinrich Schütze. 2025. On relation-specific neurons in large language models. *Preprint*, arXiv:2502.17355.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024a. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. 2024b. TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2476–2499, Bangkok, Thailand. Association for Computational Linguistics.

Viorica Marian and Ulric Neisser. 2000. Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, 129(3):361.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. 2025. On linear representations and pretraining data frequency in language models. *Preprint*, arXiv:2504.12459.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13190–13208, Singapore. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. 2024. DataTrove: large scale data processing.

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.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. *Preprint*, arXiv:2504.04264.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738, Toronto, Canada. Association for Computational Linguistics.

Yizhe Xiong, Xiansheng Chen, Xin Ye, Hui Chen, Zijia Lin, Haoran Lian, Zhenpeng Su, Jianwei Niu, and Guiguang Ding. 2024. Temporal scaling law for large language models. *Preprint*, arXiv:2404.17785.

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.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8051, Miami, Florida, USA. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2102, St. Julian’s, Malta. Association for Computational Linguistics.

A KLAR Statistics

We present the statistics of the KLAR dataset (Wang et al., 2025) in Table 2. KLAR is based on BMLAMA17 (Qi et al., 2023) with some minor modifications to improve the applicability to autoregressive models. We use **1,197** facts grouped into **12** relation categories.

Relation	**Number of Facts**
capital_of	212
continent	212
country_of_citizenship	60
headquarters_location	51
instrument	46
language_of_work_or_name	108
languages_spoken	104
manufacturer	35
native_language	130
place_of_birth	35
place_of_death	79
religion	125
total	**1,197**

Table 2: Number of facts grouped by relation types.

B Complete Factual Recall Dynamics

We present the complete factual recall dynamics in terms of *accuracy* and *crosslingual consistency* at each checkpoint of OLMo in Figure 8.

C Holistic Crosslingual Consistency

To complement the English-centric consistency analysis in the main text, we investigate **holistic crosslingual consistency**, which quantifies the agreement of correct factual predictions across **all language pairs**. Similar to §3.3, we compute the overlapping ratio of correct predictions in any two languages l and l':

$$\mathrm{CO}(l, l') = \frac{\sum_i^{|Q|} \mathbf{1}(\mathcal{M}(q_i^l) = o_i^l \wedge \mathcal{M}(q_i^{l'}) = o_i^{l'})}{\sum_i^{|Q|} \mathbf{1}(\mathcal{M}(q_i^l) = o_i^l \vee \mathcal{M}(q_i^{l'}) = o_i^{l'})}$$

where $q_i^{l'}$ and $o_i^{l'}$ are the query and expected answer for the ith query in l and l', respectively, $\mathbf{1}(\cdot)$ is the indicator function, and $\mathcal{M}(\cdot)$ is the LLM's prediction function.

We first show the crosslingual consistency between any language pairs when the model is pretrained for 400K steps. Figure 9 presents the results. We can observe that the consistency is generally low for most language pairs when the two involved languages do not share the same script, which is aligned with findings in the main text (cf. §4) that most non-Latin script languages have low consistency when compared with the predominant language, English. On the other hand, languages sharing the same script demonstrate higher similarity, for instance, Latin-script languages (fra_Latn, span_Latn, cat_Latn, tur_Latn, and eng_Latn) and Cyrillic-script languages (rus_Cyrl and ukr_Cyrl). This finding also aligns with §4, indicating that shared script has a positive effect in improving the crosslingual transfer and crosslingual consistency.

We further analyze the dynamics of crosslingual consistency within script-specific language groups, namely, Latin-script and Cyrillic-script languages, to reveal how script similarity influences consistency during pretraining. We average the consistency scores of each language pair to compute the per-group consistency. Figure 10 presents the results. We observe that consistency improves as pretraining progresses, particularly among Latin-script languages, which maintain higher mutual consistency throughout pretraining. Similarly, Cyrillic-script languages show slower but noticeable gains, but with fluctuations – possibly because only one pair of languages in this group. The overall consistency across all languages plateaus earlier. The results also align with the English-centric evaluation presented in §4. In summary, the supplementary analysis indicates that shared script and likely shared lexical structures contribute to greater alignment in factual recall across languages.

D Fact Recall and Frequencies

D.1 Overall Results

Figure 11 presents the evolution of the relationship between fact frequency and correctness across 10 checkpoints during pretraining. We observe that a linear relationship is gradually formed in the early stages (i.e., 5K to 30K steps). This linear relationship indicates that high-frequency facts are more likely to be correctly recalled than low-frequency ones. This trend stabilizes and sharpens as training progresses. This emergent frequency–correctness correlation underscores the model's bias toward memorizing frequently encountered facts. The rapid formation of this pattern indicates that pretraining quickly internalizes statistical regularities in the data, which in turn guide factual recall.

D.2 Per-Language Results

Figure 12 further breaks down the same frequency–correctness analysis by language, showing the distribution of fact frequencies and recall accuracy in each of the 12 languages. Because Dolma (Soldaini et al., 2024) is an English-centric dataset, the fact frequencies for Latin-based languages are more properly distributed. In contrast, languages of other scripts have more uneven distributions – with most facts occurring very few times or even not

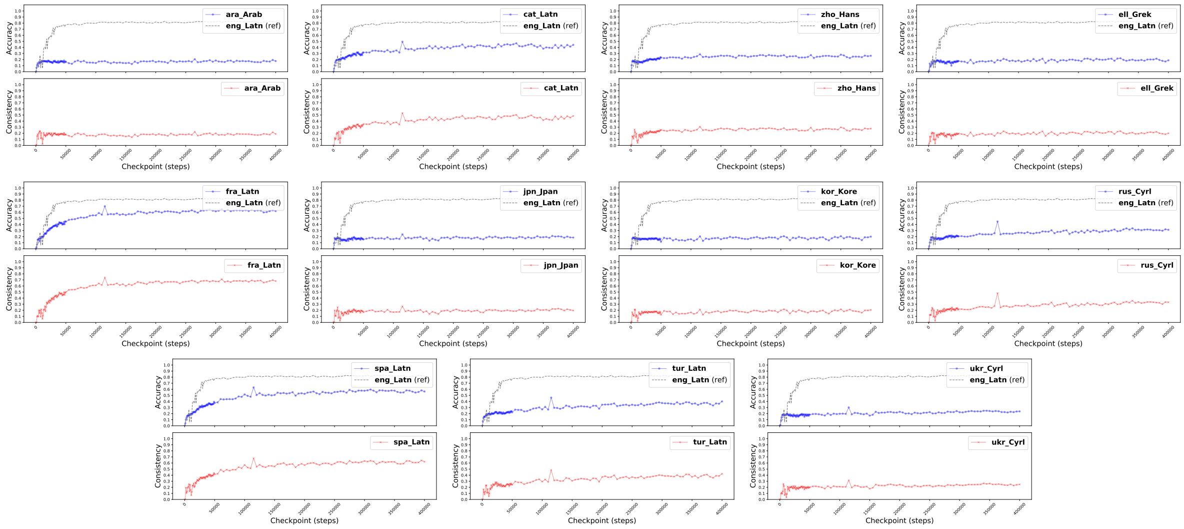

Figure 8: Factual accuracy (ACC) and crosslingual consistency (CO) for all languages.

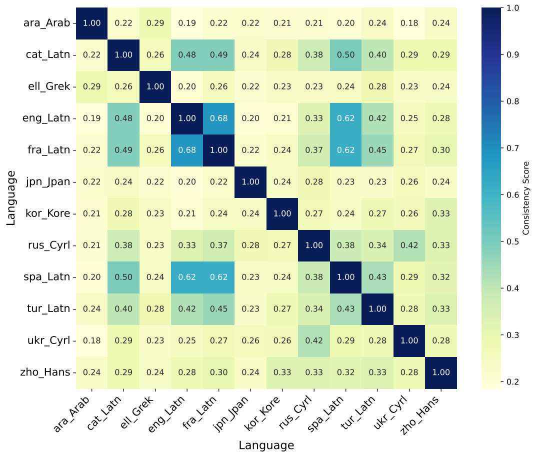

Figure 9: Crosslingual consistency of the model when it is pretrained for 400K steps. The model exhibits stronger consistency among languages that share the same script. In particular, Latin-script languages maintain consistently higher mutual consistency, while languages with distinct scripts – such as jpn_Jpan – show lower consistency with others.

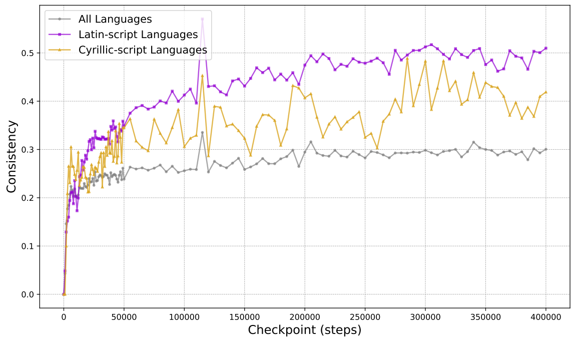

Figure 10: Dynamics of crosslingual consistency throughout pretraining. We report the average consistency among Latin-script languages, Cyrillic-script languages, and all language pairs. While consistency continues to improve among Latin-script languages and Cyrillic-script languages, the overall consistency plateaus in the early stages, which is similar to the English-centric trends observed in Figure 2.

occurring at all (not shown in the figure). However, the overall frequency–correctness correlation holds across languages, which is aligned with the global trend in §D.1. Notably, many languages have a substantial number of facts that are correctly predicted at low frequencies – mainly due to crosslingual transfer, for which we investigate in §6.

E Threshold Classifier Sensitivity

In order to analyze the sensitivity of the threshold-based classifier from Section §5.3 to the chosen threshold, we first plot the classifier accuracy for a range of thresholds within $t_l^* \pm 20\%$, for a step size of 1%, shown in Figure 13. We observe that the curves across languages are mostly flat, suggesting

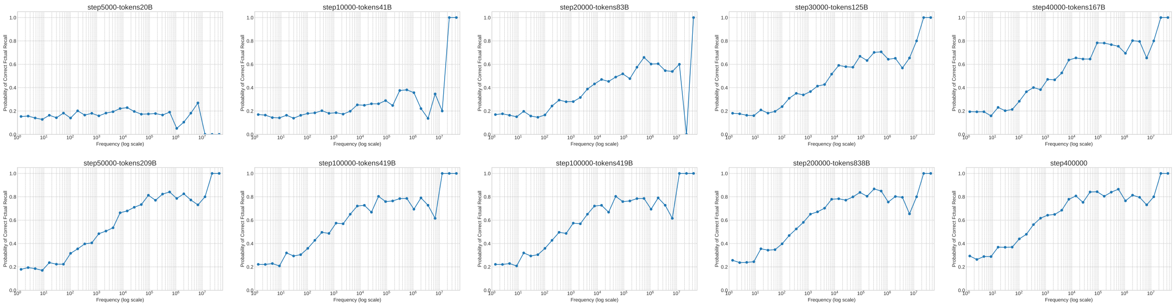

Figure 11: Relationship between fact frequency and factual recall for all languages in 10 checkpoints. High-frequency facts are more likely to be correctly recalled than rare ones. This frequency–correctness correlation emerges very early in pretraining (roughly 30K steps) and becomes more pronounced over time.

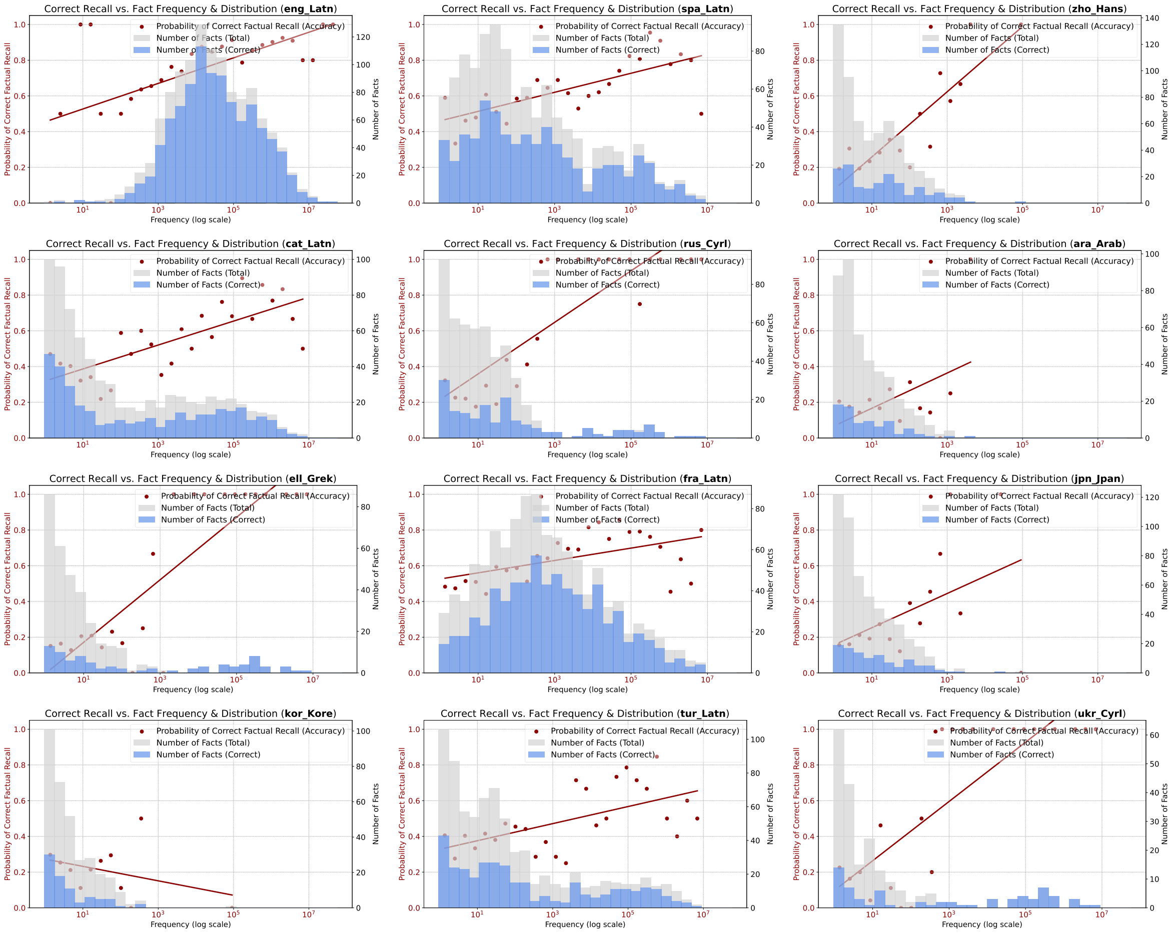

Figure 12: Complete results of the relationship between fact frequency and the probability of correct factual recall in each language.

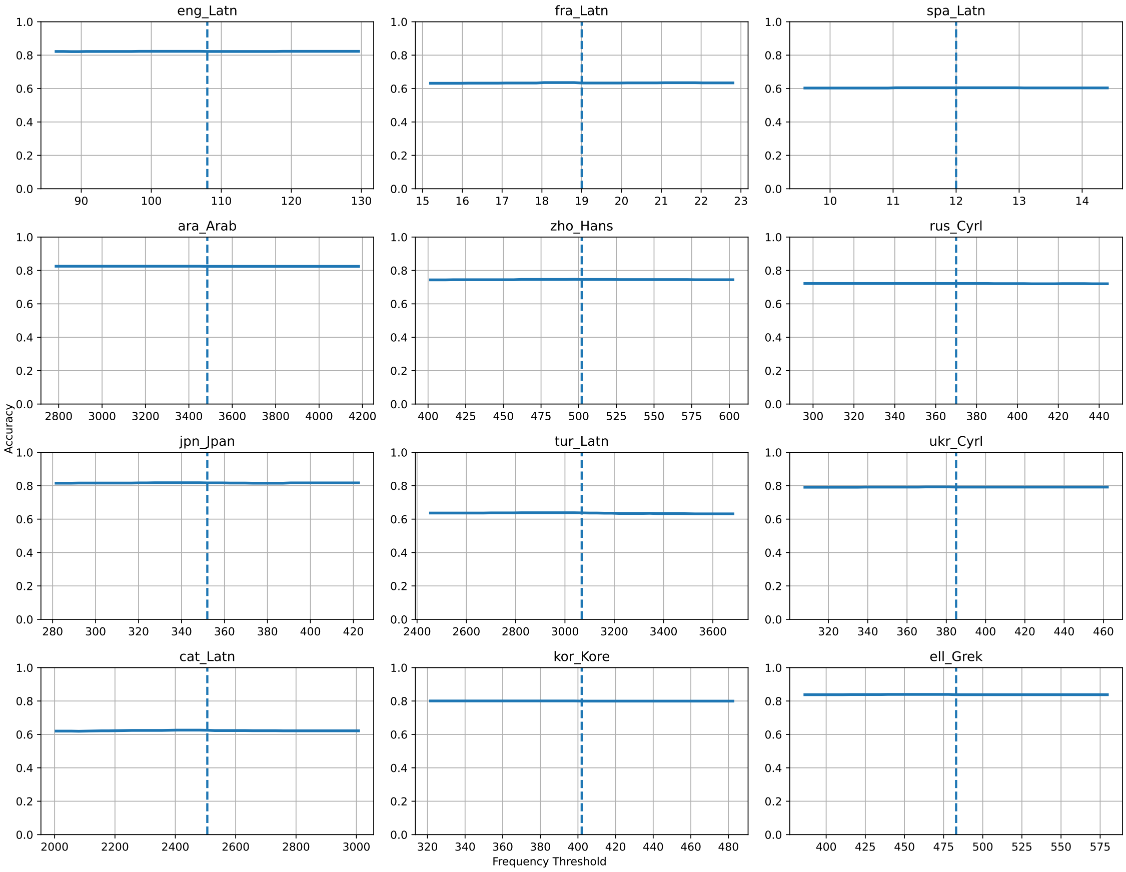

Figure 13: Classifier accuracy versus selected frequency threshold within a range of ±20% of t_l^*, the chosen threshold. The dotted line shows the actual chosen threshold.

that the classifier accuracy is robust to the chosen threshold.

To further confirm the classifier's robustness, we randomly sample 90% of the original dataset per language and select a new t_l^* based on this subsample. We evaluate the classifier on the full dataset. The results for 5000 runs are shown in Table 3. We note that though the confidence intervals for some thresholds vary widely, the resulting accuracy is very stable. Furthermore, the confidence intervals for the FP and FN counts, which are the focus of the analysis in Section §6 are narrow for most languages, with the exception of fra_Latn and spa_Latn.

We hypothesize that frequency-based prediction for these languages is confounded by two factors, both of which boost transfer from other languages: first, as we also noted in Section §5.3, fra_Latn and spa_Latn benefit strongly from transfer from English and other Latin-script languages, second, our analysis in Section §J indicates that fra_Latn and spa_Latn are well-represented in the pre-training data (cf. §6).

F Complete Similarity Progression

To supplement the representative trends shown in Figure 7, we present the full set of similarity dynamics across all 12 languages, as show in Figure 14. These plots track the mean cosine similarity between contextualized representations of fact pairs (one in English and one in the target language) across training checkpoints. We separately report trends for ***SCLFP***, ***UWLFP***, and all fact pairs, enabling a detailed view into how representation alignment evolves throughout pretraining.

Across languages and scripts, we consistently observe that ***SCLFP*** exhibit greater similarity with English than ***UWLFP***. Since both ***SCLFP*** and ***UWLFP*** are low-frequency facts, the similarity gap indicates that ***UWLFP*** are correctly recalled because their representations are better aligned with their English counterparts, while ***UWLFP*** in each language are less similar compared to the English counterparts and thus fail to benefit from crosslingual transfer. One interesting case is ukr_Cryl, where the gap between ***SCLFP*** and ***UWLFP*** is not pronounced. We hypothesize that ukr_Cryl benefits crosslingual transfer more from rus_Cryl instead of English because of shared script. The higher crosslingual consistency in the 400K-step model (cf. Figure 9) and continuously improving consistency in pretraining (cf. Figure 10) support our hypothesis. These full-language plots further strengthen our claim: pretraining on English benefits other languages not just through shared tokens or frequency-based priors, but also through crosslingual transfer from representational alignment, which goes beyond script boundaries.

G Complete Learning Dynamics on ***SCLFPs***

We present the learning trajectories of ***SCLFP***s across all languages in Figure 15.

H Complementary Analysis of Facts

To gain a deeper understanding of how factual knowledge in different languages benefits from English-centric pretraining, we conduct a complementary analysis focusing on surface-level features of facts, particularly the overlap in object strings across languages.

H.1 Same Object Effect

We hypothesize that facts in a language l that share the **same object string** as their English counterparts are more likely to benefit from transfer during pretraining. To investigate this, we report in Table 4 the proportion of facts in each language that share the same object with English, grouped by ***SCLFP*** and **non-*SCLFP*** according to our threshold-based classification (cf. §5.3).

We find that very few ***SCLFP*** share identical objects with English. This is expected since ***SCLFP*** in each language have low frequencies.[9] This finding, actually, further supports our claim that crosslingual transfer in ***SCLFP*** arises from deeper representational alignment (c.f. §6.3), not from trivial lexical overlap. In contrast, a substantial number of non-***SCLFP*** (which are mostly high-frequency facts) do share the same object string with English, especially in Latin-script languages.

To further understand the influence of object overlap, we select the subset of facts in each language whose English counterpart (i.e., same fact index) is correctly recalled by the model. These **identical-object facts** are strong candidates for crosslingual transfer from English via lexical alignment. Figure 16 shows the distribution of these facts across relation types, along with the proportion of them that are correctly recalled in each

[9] If a fact in a language has low frequency, it is very unlikely that it shares the same object with its English counterpart.

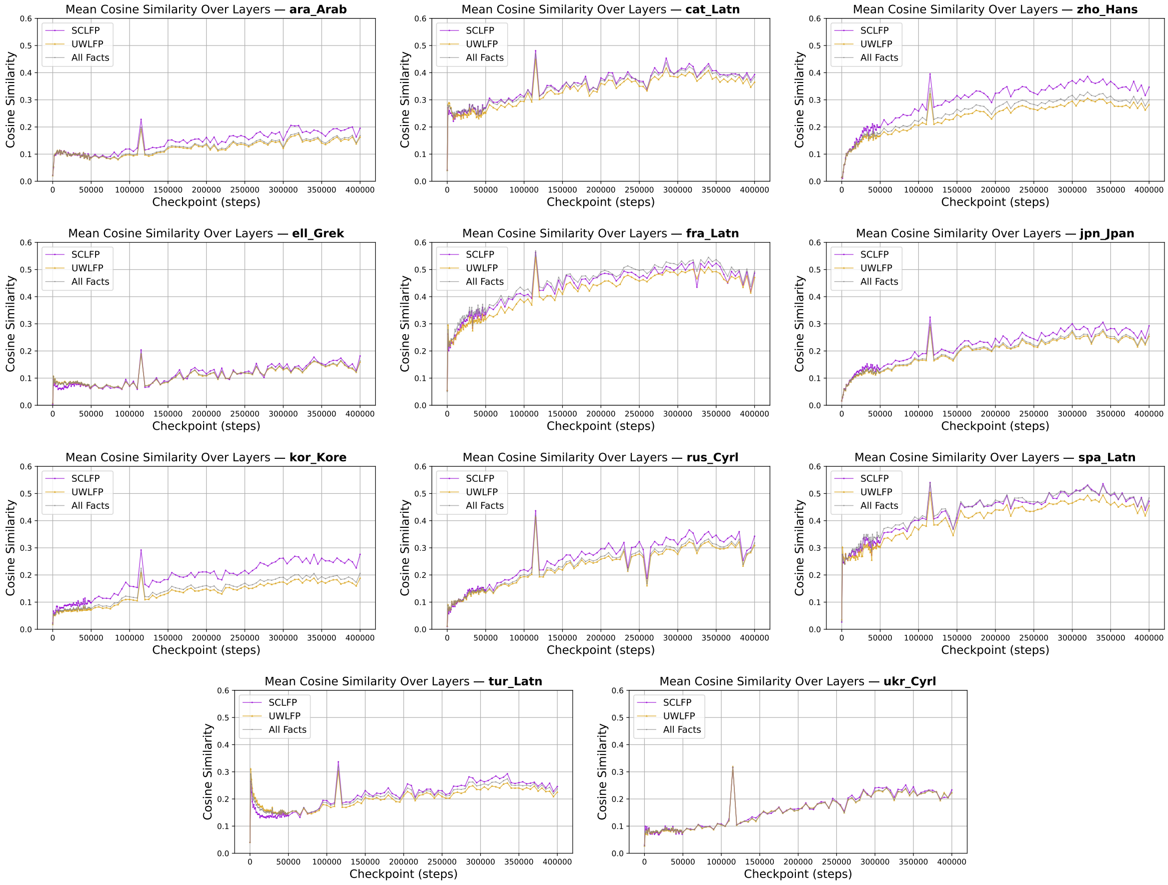

Figure 14: Complete results of mean cosine similarity for ***SCLFP***, ***UWLFP***, and all facts between each language and English during pretraining. All languages exhibit higher similarity for ***SCLFP*** compared to ***UWLFP***, indicating crosslingual transfer based on better aligned representations.

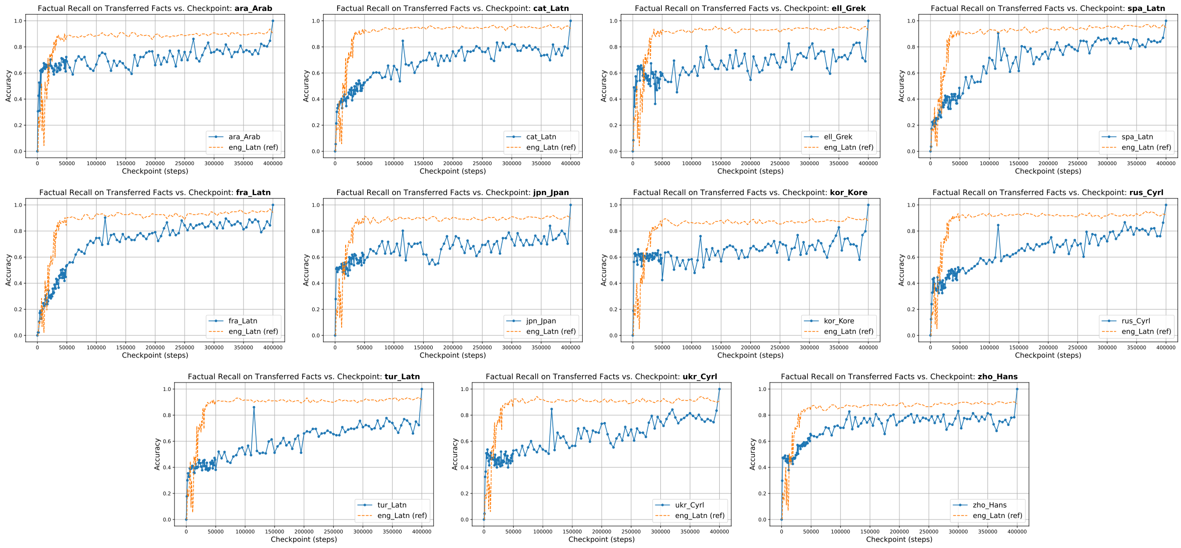

Figure 15: Dynamics of learning on the ***SCLFP***s (surprisingly correct low frequency predictions, i.e., FNs in Table 1) across all languages.

	Threshold			Accuracy			FP			FN		
Lang	**Orig.**	**Mean**	**95% CI**	**Orig.**	**Mean**	**95% CI**	**Orig.**	**Mean**	**95% CI**	**Orig.**	**Mean**	**95% CI**
ara_Arab	3485	3247	[953, 3485]	0.83	0.83	[0.82, 0.83]	0	0	[0, 3]	209	209	[208, 209]
cat_Latn	2506	2462	[2389, 2506]	0.63	0.63	[0.62, 0.63]	64	65	[64, 65]	384	383	[384, 384]
ell_Grek	483	641	[268, 1692]	0.84	0.84	[0.84, 0.84]	2	2	[0, 4]	190	190	[189, 192]
eng_Latn	108	83	[1, 146]	0.82	0.82	[0.82, 0.82]	205	207	[203, 213]	7	6	[1, 9]
fra_Latn	19	16	[5, 25]	0.64	0.64	[0.63, 0.64]	302	318	[290, 361]	134	119	[77, 146]
jpn_Jpan	352	378	[326, 450]	0.82	0.82	[0.82, 0.82]	6	6	[4, 7]	212	212	[212, 215]
kor_Kore	402	376	[262, 402]	0.80	0.80	[0.80, 0.80]	1	1	[1, 3]	238	238	[237, 238]
rus_Cyrl	370	305	[201, 370]	0.72	0.72	[0.72, 0.72]	2	4	[2, 8]	330	328	[325, 330]
spa_Latn	12	11	[5, 59]	0.60	0.60	[0.60, 0.60]	304	325	[194, 365]	169	149	[109, 287]
tur_Latn	3068	3048	[2816, 3068]	0.64	0.64	[0.64, 0.64]	60	60	[60, 61]	373	373	[373, 373]
ukr_Cyrl	385	382	[368, 385]	0.79	0.79	[0.79, 0.79]	0	0	[0, 1]	248	248	[248, 248]
zho_Hans	502	494	[461, 502]	0.75	0.75	[0.75, 0.75]	7	8	[7, 10]	296	296	[295, 296]

Table 3: Mean threshold, accuracy, false positives, false negatives, over 5000 runs of selecting a threshold t_l^* using a randomly subsampled dataset. We include the results from selecting a threshold on the full dataset for comparison, denoted "Orig.".

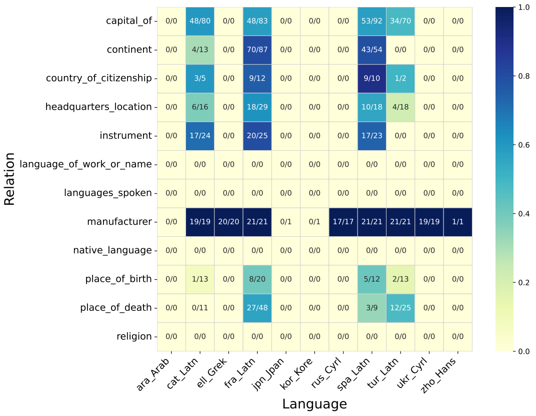

Figure 16: Distribution of **identical-object facts** across relation types for each language. A cell labeled "17/24" indicates that 17 out of 24 facts are correctly recalled, where the 24 facts are those whose English counterparts are also correctly predicted. Cells marked "0/0" indicate that no such facts exist for that relation in the given language. The results suggest that many languages, particularly those using the Latin script, benefit from sharing identical object strings with English.

language. The results confirm our expectations: Latin-script languages show consistently high recall rates for identical-object facts across multiple relation types. We also observe meaningful gains in non-Latin-script languages, particularly in the `manufacturer` relation, where object strings often reference brand names directly borrowed from English (e.g., "Apple"). These findings further highlight how both representational and lexical factors contribute to multilingual factual recall.

I Effects of Excluding Identical Facts Across Languages

In §5, we show that fact frequency can reliably predict the factual recall accuracy. The frequency of each fact is approximated by counting the number of documents where the subject and object strings of a fact co-occur. Although this measure has been widely used in previous research (Elazar et al., 2023; Merullo et al., 2025), there might be a further underlying confounding variable in the multilingual context. If two languages use the same subject/object strings for a fact, then the frequency of that fact will be the same in the two languages. This is particularly the case for Latin-script languages. For example, both French and English use "France" and "Paris", so the subject-object pair will be identical and the two languages will have the same frequency for this fact, even if sometimes the fact occurs in French text while sometimes in English text. In other words, many fact frequencies will be **aggregated statistics** over multiple script-sharing languages.[10] Therefore, we want to investigate how the results will be affected if this confounding variable is excluded.

We exclude facts in each language whose subject-object pairs match those in any other language (via string matching). This results in fewer facts in each language, but the remaining facts in each language are not affected by other languages (at least the languages considered in this study). Then we re-conduct the same investigation pre-

[10] Of course, due to the shared tokens, every occurrence of subject/object strings will affect the recallability of the fact shared by multiple languages. Therefore, we simply use the aggregated statistics for each language in the main text.

Language	**#*SCLFP***	#object matched	ratio	**#non-*SCLFP***	#object matched	ratio
tur_Latn	373	14	3.8%	824	149	**18.1%**
spa_Latn	169	6	3.6%	1028	239	**23.2%**
cat_Latn	384	11	2.9%	813	181	**22.3%**
fra_Latn	134	16	11.9%	1063	325	**30.6%**
ara_Arab	209	0	0.0%	988	0	0.0%
zho_Hans	296	0	0.0%	901	1	**0.1%**
rus_Cyrl	330	0	0.0%	867	17	**2.0%**
jpn_Jpan	212	0	0.0%	985	1	**0.1%**
ukr_Cyrl	248	0	0.0%	949	19	**2.0%**
kor_Kore	238	0	0.0%	959	1	**0.1%**
ell_Grek	190	0	0.0%	1007	20	**2.0%**

Table 4: Statistics of object agreement with English in ***SCLFP*** and **non-*SCLFP*** across languages. Many Latin-script languages tend to have a higher proportion of identical objects in **non-*SCLFP*** compared to ***SCLFP***.

sented §5.2 and §5.3.

We first present the per-language relationship between fact frequency and factual recall for **five Latin-script languages** (eng_Latn, spa_Latn, cat_Latn, fra_Latn, tur_Latn) and **two Cyrillic-script languages** (ukr_Cyrl, rus_Cyrl) in Figure 17. We observe that, even though there are fewer facts in some languages compared with Figure 12, where identical facts are not excluded, the trend still remains in each language: higher-frequency facts are more likely to be correctly predicted.

We then present the frequency-based classification for each language. Similar to the setting in §5.3, the best threshold is selected by maximizing the overall accuracy. Table 5 shows the results. We observe that there are almost no changes for languages that neither use Latin script nor Cyrillic script compared to Table 1. This is expected since only a very tiny number of facts are removed from these languages. On the other hand, we observe that there are some minor changes in Latin-script and Cyrillic-script languages. These changes are mainly in the absolute number of FP, FN, TP, TN, and Total. The best threshold has almost not changed at all except for spa_Latn, rus_Cyrl, and ukl_Cyrl, indicating the robustness of classification and similar frequency distribution before and after removing the identical facts. Since we are interested in false negatives – facts with low frequencies that are correctly predicted, we also compute the agreement between false negatives before and after the identical facts are removed. The overlapping rate is more than 98% averaged across languages, indicating that the identical facts have almost no influence on the analysis presented in the main text.

J Multilingual Coverage in Dolma

We estimate the coverage of Dolma for each language based on the frequency of token pairs. We tokenize the GlotLID Corpus (Kargaran et al., 2023), a multilingual corpus comprising texts from diverse sources, using DataTrove tokenizers (Penedo et al., 2024) specific to each language. From the tokenized output, we select the top four most frequent tokens that predominantly occur in one target language but not in the others. We then compute the frequencies of all unique, non-repetitive token pairs formed from these top tokens within the Dolma corpus. The results are presented in Figure 18. The low variance within each language's boxplot indicates that the method offers a stable and reliable comparative measure of multilingual coverage. The figure reveals a substantial disparity in pair frequency across languages, ranging from high-resource languages such as French (fra_Latn) to low-resource ones like Ukrainian (ukr_Cyrl).

K Per-Relation Dynamics Across Languages

In this section, we analyze factual recall accuracy and crosslingual consistency at the level of individual relations across languages, enabling us to examine how factual knowledge of different relation types evolves over the pretraining progression. We report the results for ara_Arab in Figure 19, cat_Latn in Figure 20, ell_Grek in Figure 21, spa_Latn in Figure 22, fra_Latn in Figure 23, jpn_Jpan in Figure 24, kor_Kore in Figure 25, rus_Cryl in Figure 26, tur_Latn in Figure 27, urk_Cryl in Figure 28, and zho_Hans in Figure 29.

Lang	Threshold	Accuracy	FP	FN	TP	TN	Total
ara_Arab	3485	0.83	0	209	1	987	1197
cat_Latn	2506	0.60	17	359	25	549	950
ell_Grek	483	0.84	2	190	4	970	1166
eng_Latn	108	0.82	156	7	740	11	914
fra_Latn	19	0.62	221	134	436	152	943
jpn_Jpan	352	0.82	5	212	8	968	1193
kor_Kore	402	0.80	0	238	1	957	1196
rus_Cyrl	201	0.70	4	319	17	744	1084
spa_Latn	5	0.59	281	106	391	155	933
tur_Latn	3068	0.64	16	369	27	645	1057
ukr_Cyrl	219	0.78	2	238	3	838	1081
zho_Hans	502	0.75	7	296	17	872	1192

Table 5: Best threshold, accuracy, and error breakdown (false positives, false negatives, true positives, and true negatives) for predicting factual recall correctness using fact frequency. For each language, we exclude facts whose subject-object pairs match those in any other language (via string matching). The results closely mirror those in Table 1, suggesting that identical subject-object facts across languages have minimal influence on the robustness of frequency predicting factual recall correctness, even for Latin-based languages and Cyrillic-based languages, which share many identical subject/objects for named entities.

We observe a similar trend as shown in §4: the consistency in each relation is primarily driven by whether the fact is correctly recalled in each language $l \neq$ eng_Latn, since the corresponding fact is almost always recalled in English.

The accuracy varies substantially across different relations within each language, with particularly large disparities in languages that use non-Latin scripts. For example, ara_Arab has nearly zero accuracy for `place_of_birth` relation.

L Experimental Environment and Hyperparameters

All experiments are conducted on NVIDIA RTX A6000 GPUs. For each fact in each language, we use the prompt template provided in KLAR (Wang et al., 2025). Each final query is accompanied by three randomly selected demonstrations to enhance pattern-matching capabilities, thereby facilitating object extraction from the model's response. We use vLLM to generate responses for each query, with generation parameters set to greedy decoding and a maximum output length of 10 tokens.[11]

[11] https://docs.vllm.ai/en/latest/

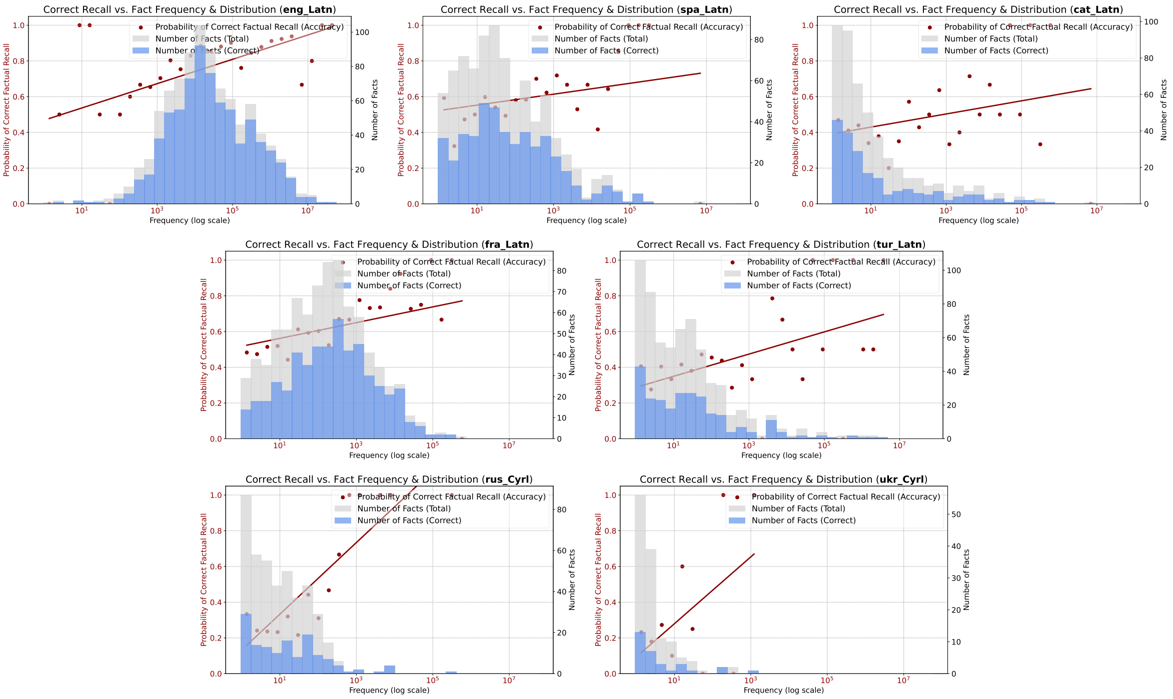

Figure 17: Relationship between fact frequency and the probability of correct factual recall for **five Latin-script languages** (eng_Latn, spa_Latn, cat_Latn, fra_Latn, tur_Latn) and **two Cyrillic-script languages** (ukr_Cyrl, rus_Cyrl) when excluding facts with subject-object pairs that exactly match those in any other languages. While shared script appears to influence the distribution of fact frequencies, a consistent trend remains across languages: higher fact frequency is associated with a higher possibility of correct factual recall.

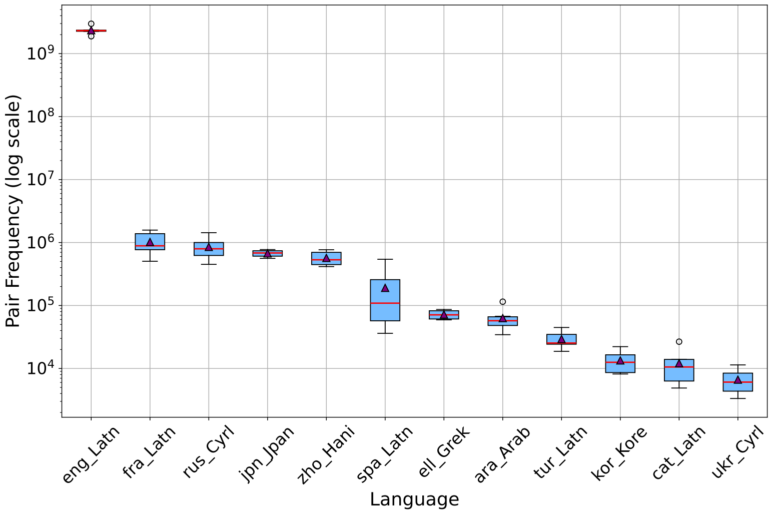

Figure 18: Pair frequency distribution (log scale) for the top four most frequent language-specific tokens in the Dolma corpus, measured across 12 languages.

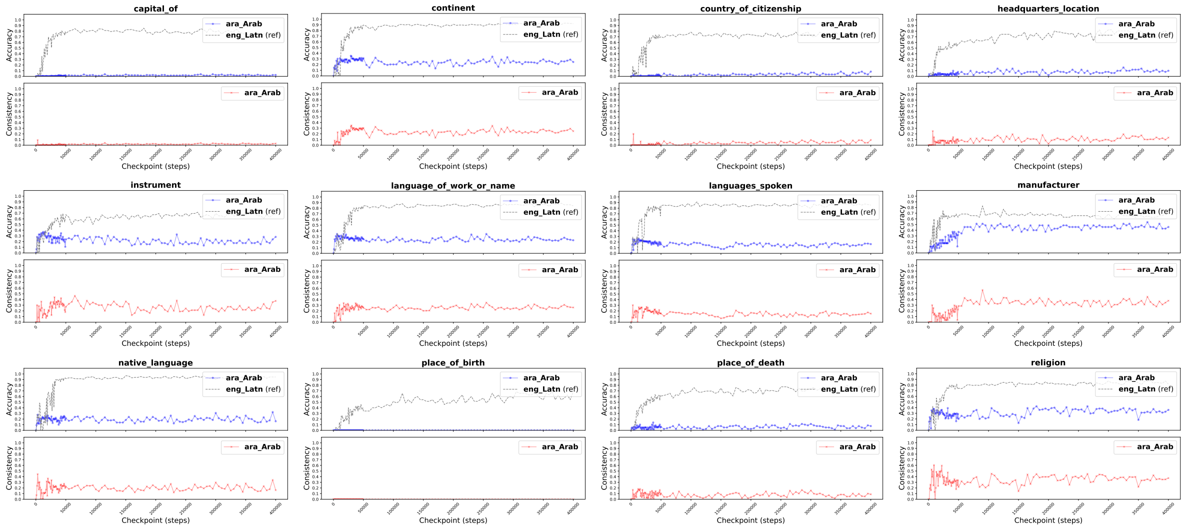

Figure 19: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **ara_Arab**.

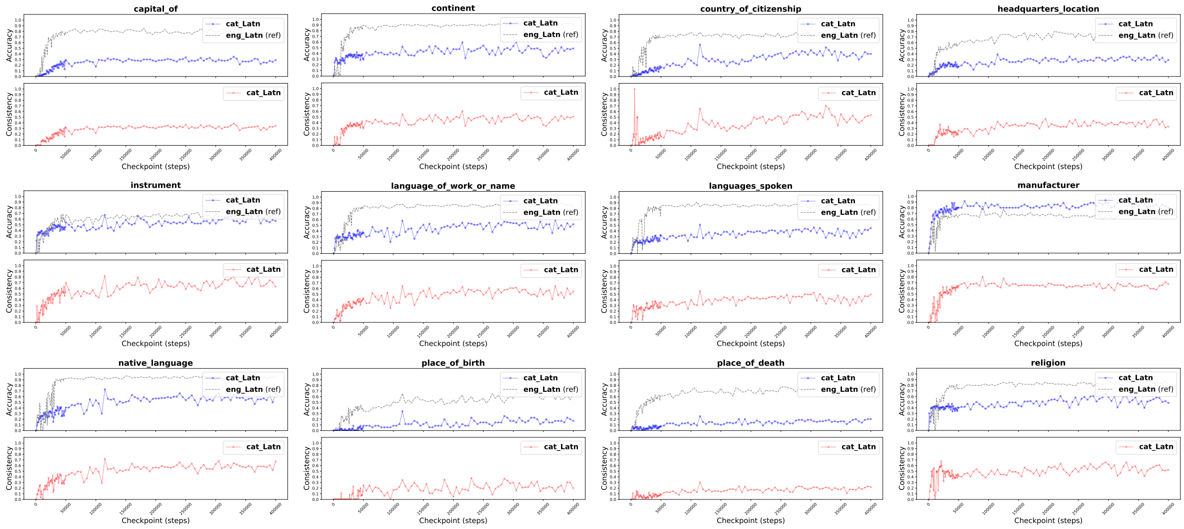

Figure 20: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **cat_Latn**.

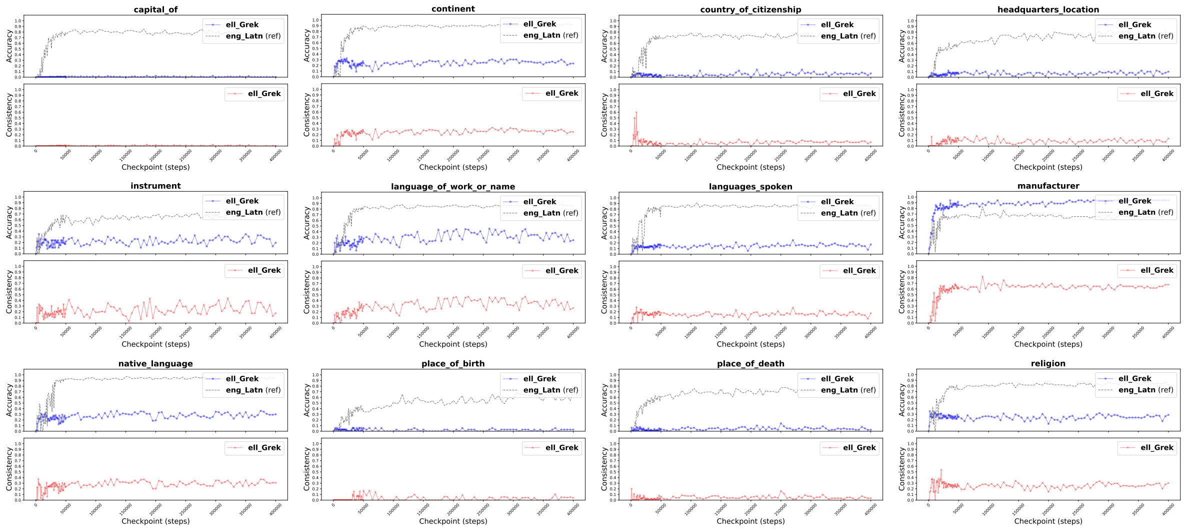

Figure 21: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **ell_Grek**.

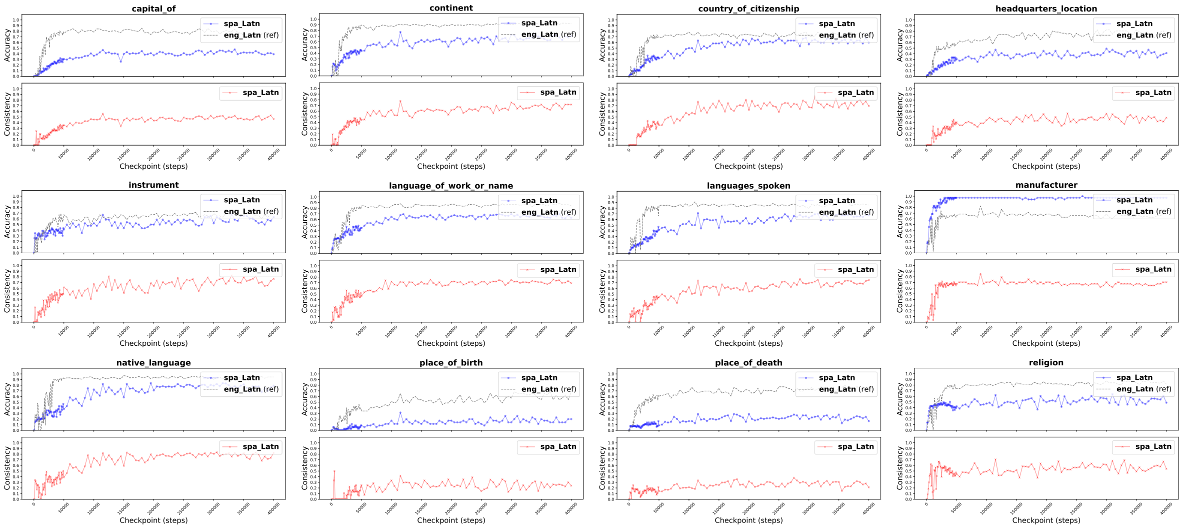

Figure 22: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **spa_Latn**.

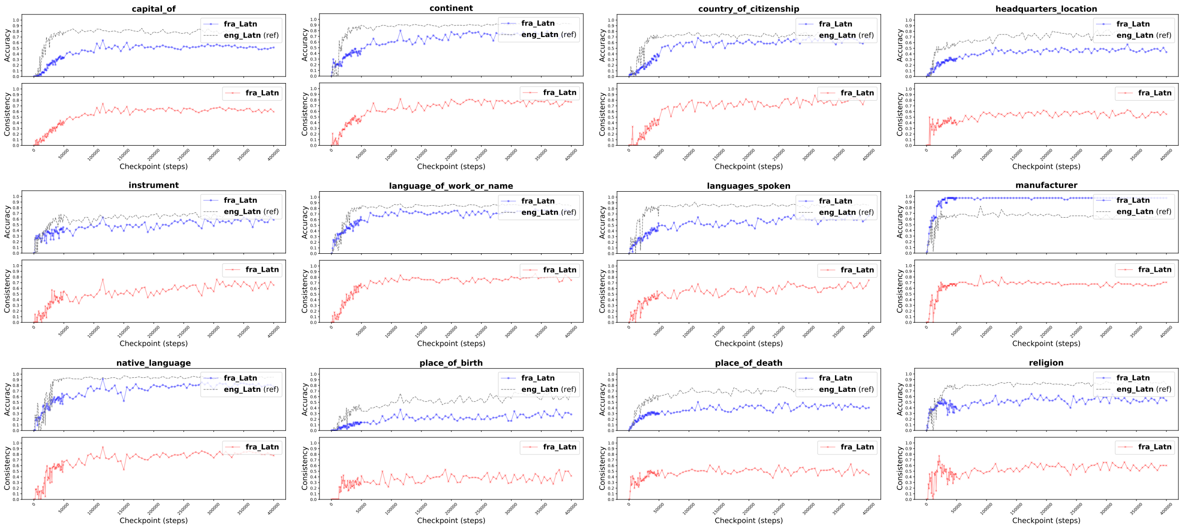

Figure 23: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **fra_Latn**.

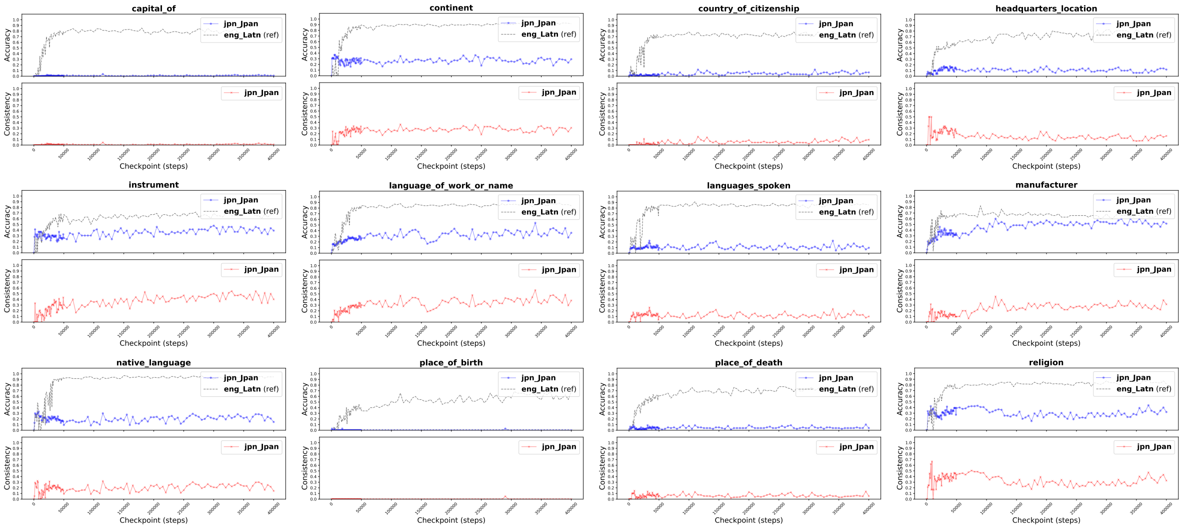

Figure 24: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **jpn_Jpan**.

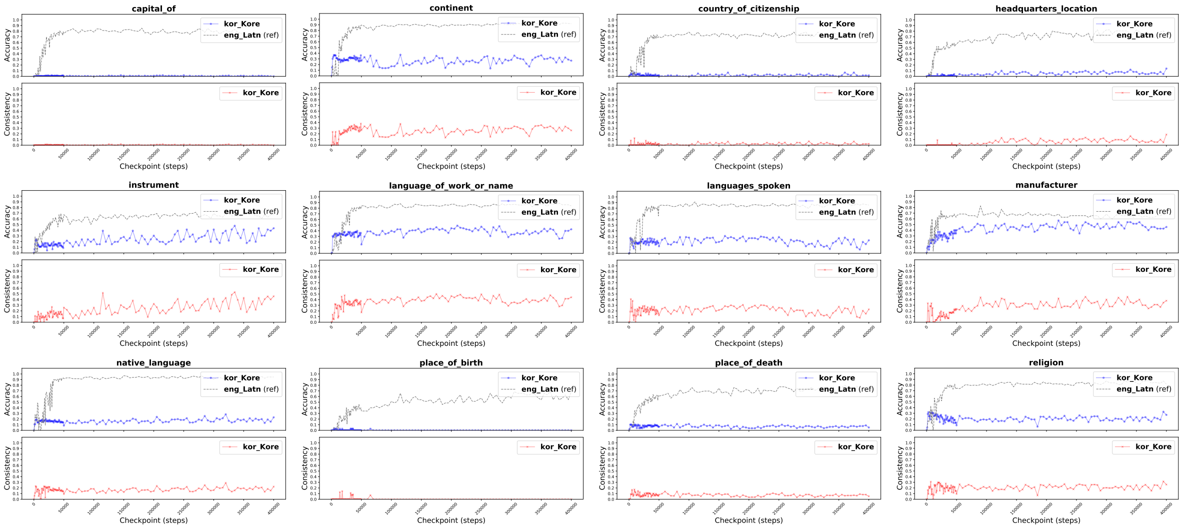

Figure 25: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **kor_Kore**.

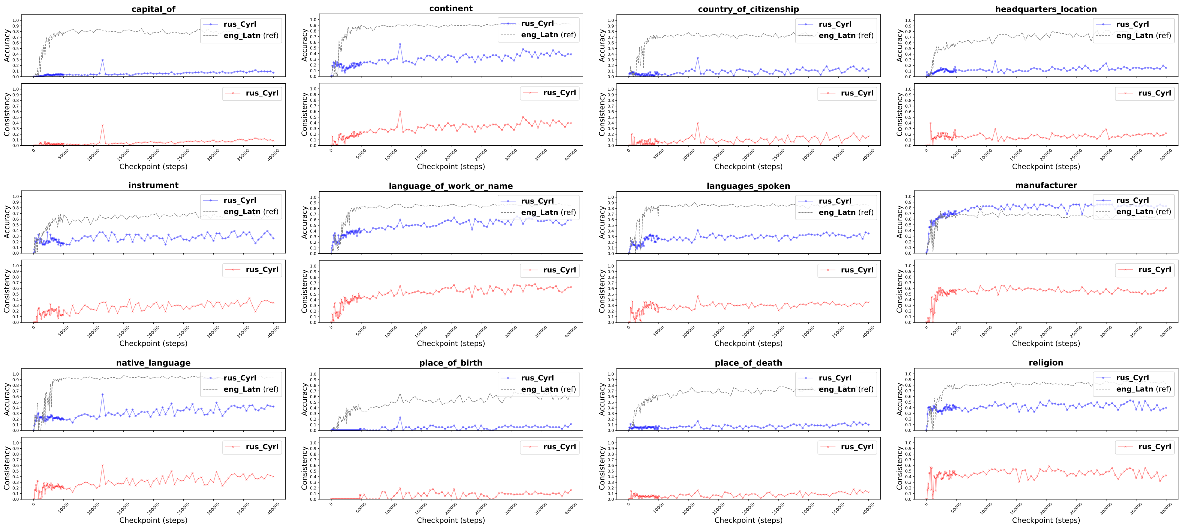

Figure 26: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **rus_Cyrl**.

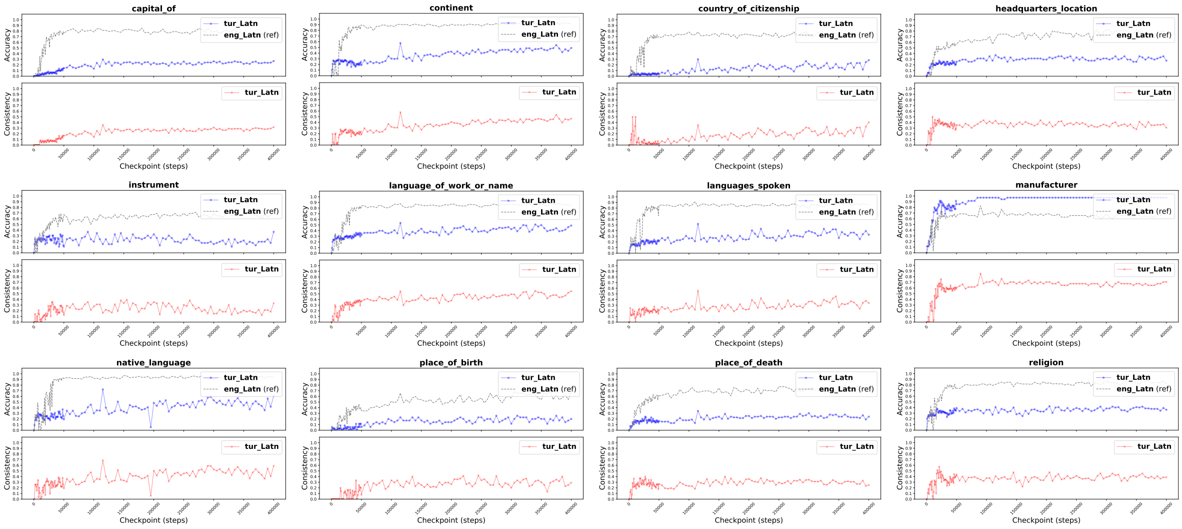

Figure 27: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **tur_Latn**.

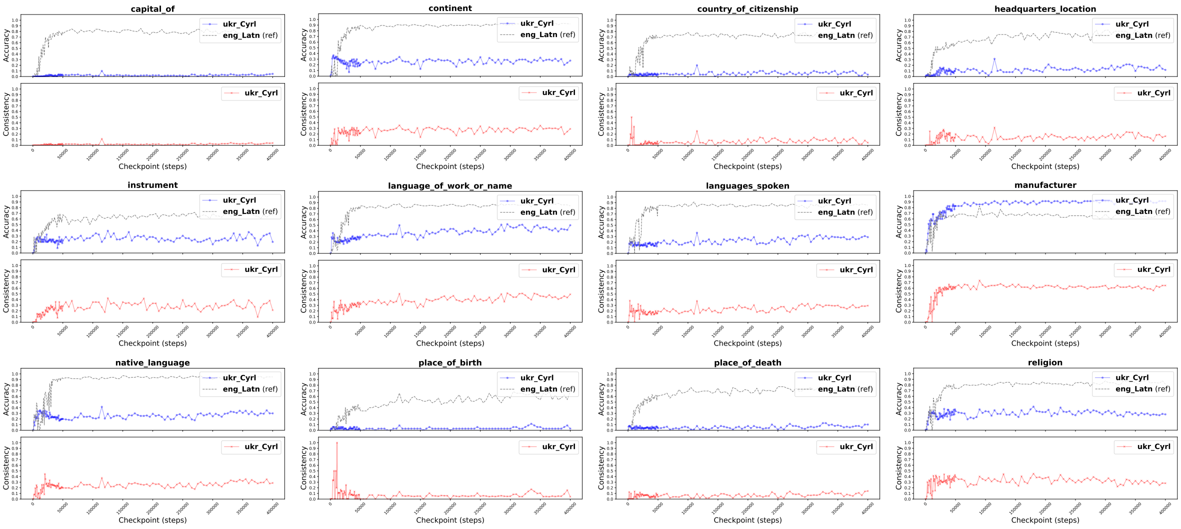

Figure 28: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **ukr_Cyrl**.

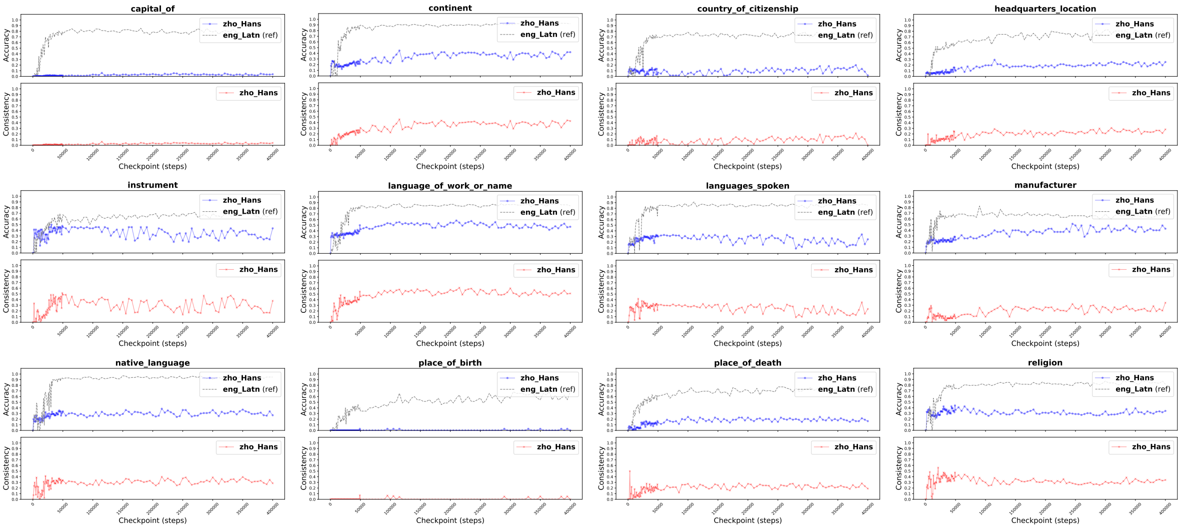

Figure 29: Factual accuracy (ACC) and crosslingual consistency (CO) for each relation type in **zho_Hans**.