WHY PRETRAINING FAILS TO SHARE CROSS-LINGUAL KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have made remarkable progress in the processing and modeling of many languages. Yet, unlike human multilinguals, they exhibit surprisingly limited cross-lingual knowledge transfer. While this limitation is well documented, its origins during multilingual training remain unclear. We pretrain 360M- and 7B-parameter LLMs and show that poor cross-lingual knowledge generalization emerges during pretraining and persists under standard interventions. To isolate its cause, we employ a controlled bilingual pretraining setting using two copies of the same language, sharing identical text and token segmentation, but mapped to disjoint token spaces. We find that disjoint tokens alone are enough to induce knowledge compartmentalization, even between identical copies of the same language, establishing disjoint token spaces as a fundamental barrier to cross-lingual knowledge generalization. Guided by this understanding, we suggest mapping languages into a shared token space by simple word-wise translation and find it substantially improves cross-lingual knowledge generalization, recovering up to 12.6% of native-language learning efficiency — 14× the baseline.[1]

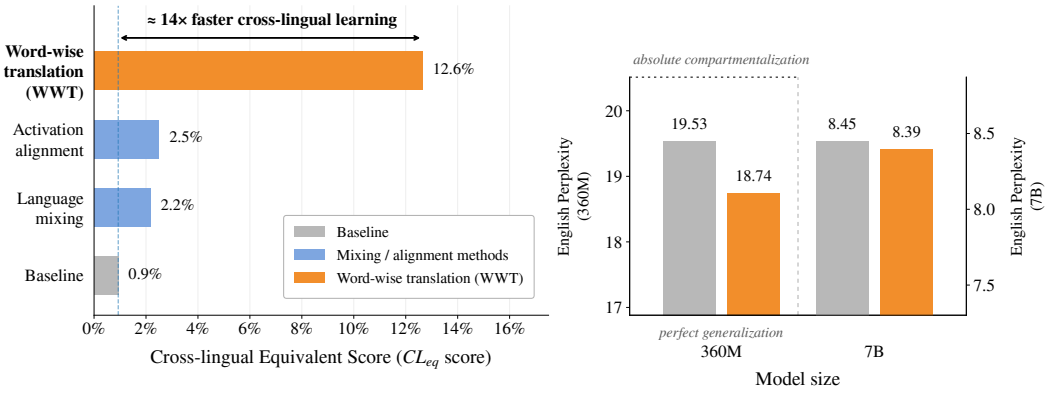

Figure 1: Merging token spaces via WWT yields substantial gains in cross-lingual knowledge generalization. Left: CL_{eq} for English–Arabic pretraining under various methods (colors), measuring cross-lingual knowledge learning efficiency at 360M scale. Right: left panel shows English perplexity at 360M; upper bound is a model trained on the English half alone, the lower bound is a model trained on twice as much English. Right panel shows perplexity at 7B, bounds are proportional to 360M.

1 INTRODUCTION

Modern LLMs exhibit impressive multilingual capabilities, yet unlike humans, they transfer surprisingly little knowledge across languages. Understanding this limitation is important for building truly multilingual models.

Prior work has documented poor cross-lingual generalization, characterizing it through model capacity and sample efficiency, and exploring factors such as tokenization, language imbalance, and activation

[1] Code publicly available https://anonymous.4open.science/r/torchtitan-mulitlingual-D661/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

alignment (Chang et al., 2024b; Schäfer et al., 2024; Howe & Wingate, 2026; Wu & Dredze, 2020; Li et al., 2024b). Recent work extends this generalization failure into the knowledge domain (Qi et al., 2023; Ifergan et al., 2025; Goldman et al., 2025) and tries to reduce it (Huang et al., 2023; Diskind et al., 2026; Bandarkar et al., 2026; von Rad et al., 2026a). However, these observations and fixes are made post-hoc on fully trained models, which go through a sequence of training stages including pretraining, instruction tuning, and preference optimization, leaving open where this knowledge generalization failure, which we refer to as compartmentalization, originates.

We investigate compartmentalization directly at the pretraining stage. However, causal analysis in this setting is complicated by two factors: First, pretraining corpora are vast, and tracking knowledge exposure across languages is intractable. Second, natural languages simultaneously differ in vocabulary, scripts, syntax, frequency distributions and semantics, hindering controlled experiments.

We address these challenges through a two-part methodology. First, to reliably trace knowledge exposure, we introduce a controlled pretraining environment utilizing the injection of fictive knowledge (§2). Second, to strip away linguistic confounds, we evaluate generalization not only on natural bilingual pairs but also on structurally identical, cloned language copies, that differ in nothing but their token spaces (§5.1). Using this combined framework, we pretrain 360M- and 7B-parameter models to measure and isolate the mechanisms of cross-lingual knowledge generalization.

Our controlled measurements reveal that knowledge **compartmentalization already emerges during pretraining** (§4). We further identify **disjoint token spaces as a fundamental barrier** to the knowledge generalization mechanism that persists across model scales. Two languages represented in disjoint regions of the model's vocabulary, regardless of token segmentation, hinder sharing (§5).

Guided by our newfound understanding of this mechanism, we suggest a word-wise translation (WWT) intervention to efficiently unify the token space of vocabulary-disjoint languages (§6.1). This intervention proves beneficial, substantially improving cross-lingual knowledge generalization while providing gains in general language modeling perplexity (Fig. 1). This practical success provides a mechanistic validation of our findings on tokenization's role in knowledge generalization (§6.2).

2 Methods: Measuring Knowledge Generalization

2.1 Controlling knowledge exposure - Injecting fictive facts

To measure if and to what extent exposure to a fact during training in one language yields knowledge of that fact in another language, we must know how often the fact has been observed in each language. Tracing fact exposures in the vast pretraining corpora is often intractable (Kandpal et al., 2023; Elazar et al., 2024). We bypass this problem by injecting novel fictive knowledge, absent from the underlying pretraining data, with precise control over the number and language of exposures to each fact.

Our Fictional Knowledge Dataset (FKD) contains 2,048 facts about synthetic entities and associated attributes (e.g., "Cultramive is a company that sells silver-plated bicycles"). Each fact is expanded into a per-language set of diverse natural documents that paraphrase the same underlying knowledge — used for injection during pretraining — and held-out multiple-choice questions (MCQs) for evaluation. We provide full details on FKD, including the generation pipeline and examples in App. C. Since prior work indicates that cross-lingual generalization mechanisms can behave differently for high- and low-frequency tokens (Schäfer et al., 2024; Feng et al., 2025), we replicate our main experiments on a fictive-persona dataset built around high-frequency human names (App. D).

We validate FKD by pretraining monolingual models with entity facts injected at different rates (App. C.5). At 0-injection, accuracy on the MCQs is around chance, injecting fact paraphrasings into pretraining increases accuracy in a log-like manner, reaching substantial accuracy across all languages; FKD is learnable and controllable.

2.2 Evaluating Cross-lingual knowledge generalization – CL_{EQ} Score

While fictive knowledge allows us to control how it is introduced across languages, we're still missing a direct measure, given a bilingual model, that quantifies how well it generalizes factual knowledge between languages A and B. To this end we introduce the Cross-Lingual equivalence (CL_{eq}) score.

$\text{CL}_{\text{eq}}{}_{B\to A}$ measures how valuable an exposure to a fact in language B is for recalling that fact in language A, relative to a native exposure in A itself. A score of 100% indicates that exposure in B is as effective as exposure in A, corresponding to perfect cross-lingual equivalence. A score of 0%, in contrast, indicates that exposure in B provides no benefit for recall in A, corresponding to complete knowledge compartmentalization across the two languages.

When reporting a single score for a given bilingual model trained on languages A and B, we use the bidirectional average:

$$\text{CL}_{\text{eq}} = \frac{1}{2}\left(\text{CL}_{\text{eq}}{}_{A\to B} + \text{CL}_{\text{eq}}{}_{B\to A}\right).$$

Crucially, our CL_{eq} definition normalizes cross-lingual gains by native ones. This makes our metric invariant to differences in baseline model competence (raw accuracy) that may arise from scale, training recipe, or data quality; a model that learns all facts better raises foreign and native exposure value together, leaving their ratio unchanged. This separates gains in target-language ability from gains in transfer, which measurements reported in absolute accuracy terms conflate.

Estimating CL_{eq} score Controlling FKD exposure enables a direct estimate of a model's CL_{eq} score. For a fictive fact f in FKD, let $N_A(f)$ and $N_B(f)$ denote the number of exposures to it during pretraining in languages A and B respectively, and let $\text{acc}_A(f)$ denote trained model accuracy on f's language-A evaluation questions. We fit a linear regression (OLS) over all facts,

$$\text{acc}_A(f) = \alpha + \beta_A\, N_A(f) + \beta_B\, N_B(f) + \varepsilon, \tag{1}$$

β_A captures the accuracy gained per native exposure, β_B captures the accuracy gained per cross-lingual exposure, and the score is derived from their ratio:

$$\text{CL}_{\text{eq}}{}_{B\to A} = 100 \times \frac{\beta_B}{\beta_A}. \tag{2}$$

Two features of this estimate warrant note. First, the OLS regression provides a linear approximation to the relationship between exposure and accuracy (§2.1); we experiment with an alternative curve-aware estimator and find the simple linear form more informative. Second, sampling and evaluation noise can occasionally pull estimates slightly below zero. Negative values should be read as substantial or absolute compartmentalization rather than negative transfer. App. E covers these topics in detail, and reports significance measures confirming our headline comparisons are statistically robust.

3 Experimental Setup

Our experiments follow a common template. We pretrain bilingual models from scratch, injecting FKD fictive facts at varying cross-lingual rates, and estimate CL_{eq} score based on their acquisition. We vary isolated aspects of the bilingual setup (architecture, tokenization, and interventions) and compare the resulting scores to reason about cross-lingual knowledge generalization. Each of the presented CL_{eq} results requires a full model pretraining run corresponding to its unique condition.

3.1 Models and Data

We evaluate two scales: a 360M-parameter model following the SmolLM2 architecture (Allal et al., 2025), and a 7B-parameter adaptation of Llama-3 (Grattafiori et al., 2024). Both use a BPE tokenizer with a 65,536-token vocabulary. We train one tokenizer per language pair on a 50/50 split of 7M documents from each of its two languages, and share it across the matching runs. Models whose perplexities we compare therefore read text under identical segmentation. Models are trained to approximately Chinchilla-optimal token budgets (Hoffmann et al., 2022) using TorchTitan (Liang et al., 2025). Architectures, hyperparameters, schedules, and hardware are detailed in App. B.

We pretrain on *FineWeb-edu* for English (Penedo et al., 2024), *FineWeb-edu-ar* for Arabic (Alrashed et al., 2024), a machine translation of FineWeb-edu, and *FineWeb2-HQ* for Russian (Messmer et al., 2025), native web text rather than translation. Our bilingual experiments use a 50/50 token budget split between languages. As the English and Arabic corpora are translations of one another, we enforce a strict document-disjoint split to ensure models do not see parallel documents across languages.

3.2 EVALUATION

To calculate the CL_{eq} metric, we split FKD into 16 groups of 128 fictive entity facts. Each of the 16 groups has all its fictive facts injected at a uniform N_A and N_B chosen from the cross-product of [0,20,100,1000] with itself. This range was selected to be denser on the critical ranges of the fact learning curve (App. C.5). Fact paraphrasing documents are injected stochastically into pretraining at a per-language probability matching the target exposure counts N_A and N_B (App.B.5).

We use the LM-eval-harness (Gao et al., 2024) evaluation protocol along with the accompanying FKD per-language evaluation MCQs (generated separately from the injection documents and never present in training) to quantify the model's fact acquisition across the different languages: acc_A and acc_B. Specifically, we use acc-norm scoring from LM-eval-harness to rectify for possible variance in choice length. With these values at hand, we fit the regression of Eq. 1 across all facts and take CL_{eq} as the ratio of β_A and β_B (§2.2). We report fitted coefficients for representative runs in App. E

4 KNOWLEDGE COMPARTMENTALIZATION ORIGINATES IN PRETRAINING

4.1 COMPARTMENTALIZATION EMERGES IN STANDARD BILINGUAL PRETRAINING

Applying the above-mentioned tools, we show that current pretraining practices strongly compartmentalize knowledge. We pretrain a 360M-parameter English–Arabic LLM, injecting fictive facts at controlled cross-lingual rates (§2.1). With the trained bilingual model at hand, we measure fictive fact recall and deduce the relative value of foreign exposure to knowledge — the CL_{eq} score (§2.2). The pretrained English–Arabic model shows minuscule knowledge generalization, with a CL_{eq} score of 0.9% (Fig. 1). Severe knowledge compartmentalization is also apparent in an English–Russian setup with $\text{CL}_{\text{eq}} = 1.9\%$ and at the 7B parameter scale with $\text{CL}_{\text{eq}} = 5.9\%$ (Tab. 2), suggesting knowledge compartmentalization in pretraining is not an artifact of our specific language pairing or scale.

Analyzing per-language perplexity, we find that a bilingual setup is slightly better than pretraining on either of the monolingual halves of the corpus alone (Fig. 1), consistent with previous works demonstrating positive transfer (Conneau et al., 2020a). However, it appears that whatever is driving this gain in language modeling, it is not knowledge generalization.

4.2 STANDARD INTERVENTIONS FAIL

As bilingual pretraining naturally compartmentalizes knowledge, we experiment with pretraining interventions suggested for related purposes to resolve it, and find that they largely fail. In particular, we experiment with two types of pretraining interventions in the English–Arabic setting at 360M-parameter scale: language mixing and activation alignment (App. F).

Language mixing introduces multiple languages within a single input sequence, encouraging the model to utilize cross-lingual information. Our language mixing variant of choice is pretraining code-switching (named after linguistic code-switching), which randomly substitutes words with their cross-lingual counterparts (e.g. Zeng et al., 2026). We use a comprehensive word-level dictionary for the substitution operation, and sweep both the fraction of mixed documents and the within-document substitution rate, finding no substantial benefit, with a CL_{eq} score of 2.2%.

Activation alignment uses auxiliary losses to align pooled activations of semantically equivalent cross-lingual units. We experiment with InfoNCE (van den Oord et al., 2018; Chi et al., 2021) and L2 loss estimating sequence matches by chunking parallel documents. Across various chunk sizes we find activation alignment offers no help in resolving compartmentalization: CL_{eq} score of 2.5%.

Taken together, our results establish that knowledge compartmentalization emerges already in pretraining, persists under standard interventions, and is attenuated but not resolved by scale, suggesting that a fundamental solution should address the breakdown at this stage. What causes this breakdown remains open: natural language pairs differ in script, grammar, segmentation, token vocabulary, and data distribution at once, so these measurements alone cannot say which difference is responsible.

5 Disjoint Token Space is a Root Cause of Compartmentalization

In this section, we isolate disjoint token spaces from other confounding variables of multilingual settings using a clone-languages setup (§5.1), and find that it alone is sufficient to induce knowledge compartmentalization in pretraining (§5.2). We quantify the amount of token space sharing needed to resolve it (§5.3), and show that even clone-languages shared initialization surprisingly struggles, although to a much lesser extent, to completely resolve compartmentalization (§5.4). This section's results are presented in Figure 2.

5.1 Isolating disjoint tokenization: English$_1$–English$_2$ setup

Poor multilingual generalization is frequently attributed to varying syntax, grammar, cultural nuances, data quantity and quality, or tokenizer fragmentation (Rust et al., 2021; Petrov et al., 2023; Philippy et al., 2023; Lauscher et al., 2020; Hershcovich et al., 2022; Li et al., 2024a). To strip away these confounding variables and isolate the pretraining process from linguistic aspects, we train on two copies of one, resource-rich language, rather than on two different natural languages, which we name the English$_1$–English$_2$ pretraining setup. By duplicating the model's vocabulary, we create two structurally identical languages mapped to entirely disjoint token spaces (K et al., 2020).

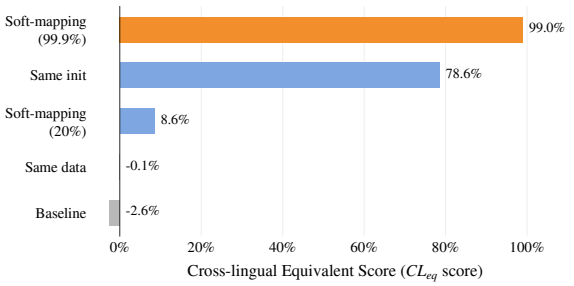

Figure 2: Disjoint token spaces are severely compartmentalized. CL_{eq} for English$_1$–English$_2$ pretraining under various methods. Soft-mapping percentages refer to the tied embedding portion p. "Same data" refers to pretraining the language copies on identical corpora

Formally, assuming a standard token vocabulary of size V, any text is tokenized into a base sequence $T = [t_1, t_2, \ldots, t_n]$ where each token index $t_i \in \{0, \ldots, V-1\}$. We define the representations for our two language copies as perfectly parallel sequences shifted into mutually exclusive token spaces:

$$\begin{aligned} \text{English}_1(T) &= [t_1, t_2, \ldots, t_n] \quad t_i \in [0, V-1] \\ \text{English}_2(T) &= [t'_1, t'_2, \ldots, t'_n] \quad t'_i = t_i + V \in [V, 2V-1] \end{aligned} \tag{3}$$

We double our model vocabulary to accommodate the disjoint token spaces. Under this setup, the sole barrier to cross-lingual generalization is the ability of the training process to bridge disjoint tokens.

5.2 Disjoint tokens alone suffice for compartmentalization

Despite an identical underlying structure, we find severe knowledge compartmentalization between English$_1$ and English$_2$: CL_{eq} score of -2.6%. Scaling up doesn't help, with a 7B-parameter model trained in this setting yielding -1.5%. Even pretraining on identical corpora (except for the injected FKD a tiny fraction of training tokens) does not offer any help (-0.1%); pretraining fails to generalize knowledge across 2 identical data distributions based on disjoint tokens alone.

We analyze the learned embedding spaces of English$_1$ and English$_2$ and find they are structurally similar and substantially alignable by a linear transformation, yet with no correspondence between matching tokens. The model successfully learns the same underlying distribution independently in each space, yet fails to bridge them. In contrast, models that do generalize learn to bridge the two token spaces, so that matching tokens have similar embeddings (App. I). Together, this shows that structural similarity is not sufficient for knowledge generalization, which requires the model to learn to bridge the two token spaces, and that the standard pretraining process does not do so on its own.

5.3 How much sharing is needed?

Disjoint Tokenization forms the barrier; we now quantify the barrier's size when tokenization is only partially disjoint. For this we introduce *soft-mapping*, where for each token pair t_i, t'_i a fraction p of their embedding is tied and learned together.

Soft-mapping experiments reveal a sharp threshold. At 20% tying, generalization remains in the weak region with $CL_{eq} = 8.6\%$; at 50% tying it jumps to 82.7%, and approaches 100% as tying becomes complete (App. G).

These results imply that a large amount of token sharing is required to induce knowledge generalization in pretraining. We also experiment with vocabulary overlap, a subset of tokens completely shared between the 2 languages, and find that effectiveness is also dependent on the frequency of the shared tokens (App. J). As a side effect, the results also confirm that our metric's ceiling is reachable, so substantial shortfalls below it reflect genuine knowledge compartmentalization.

5.4 Pretraining drifts aligned tokens apart

Token sharing is needed, but pretraining does not establish it on its own; we now test if pretraining can maintain and utilize a supplied matching. We initialize each pair of parallel embeddings in English_1–English_2 to identical values — the same alignment that complete tying enforces (§5.3), but provided as a starting point rather than a structural constraint. We find that the training process fails to utilize these matches to their full extent, yielding a CL_{eq} score of 78.6% — high, but substantially short of the reachable ceiling. Identical data exposure remains short, with a score of 84.6%. This compounds the flaw established above: pretraining not only struggles to bridge disjoint tokens on its own, it also actively drives them apart when a matching is already provided. Alignment between disjoint token spaces is thus not a stable state of pretraining but one it works against, rendering initialization-based solutions to compartmentalization incomplete on their own.

Together, this section's results establish disjoint token spaces as a fundamental barrier to knowledge generalization. Pretraining neither converges to token sharing on its own, nor preserves it when supplied. These findings motivate bypassing disjoint tokenization directly rather than aligning across it, but two questions remain. Can token spaces be merged efficiently and practically? And while disjoint tokens are sufficient to cause compartmentalization, is removing them sufficient to resolve it in realistic settings, where many other variables are in play?

6 Unifying Token Spaces Reduces Compartmentalization

Having established that disjoint token spaces are sufficient to induce knowledge compartmentalization, we now extend our mechanism diagnosis by introducing an efficient vocabulary mapping technique bypassing disjoint token spaces directly (§6.1) and demonstrating it substantially improves generalization in the natural bilingual setting (§6.2). We also verify the choice of a semantically meaningful mapping (§6.4) and address practical aspects of this mapping strategy (§6.3).

6.1 Unifying token spaces via Word-Wise Translation

The WWT operation. Word-wise translation (WWT) maps text in language A to the token space of language B by replacing each word with its counterpart from a large-scale invertible dictionary; out-of-dictionary words are transliterated via an invertible character-level mapping (dictionary curation and conflict resolution in App. H). Unlike full document translation, this mapping is context-free and word-wise, introducing minimal computational overhead. It therefore can be applied and inverted on the fly during both pretraining and inference; users read and write native-script text while the model trains and operates in the unified token space, making the intervention transparent at the interface.

WWT-Ar. We instantiate WWT for Arabic into English token space using an invertible Arabic-to-English dictionary, with the invertible Buckwalter transliteration as the out-of-dictionary fallback. The result is WWT-Ar — a new writing system for Arabic that retains every word choice, word order, and grammatical structure, while sharing a unified token space with English (Table 1). The dictionary covers 99.7% of word occurrences in our corpus ($\sim$78% of unique word types), so the mapped tokens are overwhelmingly translations rather than transliterations. We also apply WWT to Russian, unifying its token space with English and creating WWT-Ru.

Cost. WWT increases sequence length: WWT-Ar requires roughly 30% more tokens than native-script Arabic to encode the same text (App. H.3), increasing the cost of LLM inference. Under the fixed token budget of our experiments, this fertility overhead also means that a WWT-Ar model is exposed to roughly 23% fewer Arabic documents than its native-script counterpart, resulting in a data

disadvantage relative to the models we compare it against. We leave the overhead unhandled and address its origins and paths to reducing it in App. A.

6.2 Unified token space improves knowledge generalization

Language	Text
Arabic (original)	فشلت اللغات المنفصلة في مشاركة المعرفة←
WWT-Ar	failed the languages the disjoint in sharing the knowledge →
English (translation)	The disjoint languages failed to share knowledge →

Table 1: An Arabic sentence, its WWT-Ar rendering, and an English translation. Arrows signal reading direction, color marks word origin. WWT-Ar maps every word into English tokens while preserving Arabic word order and grammar; it is Arabic in a different writing system, not a translation.

		CL_{eq}		English PPL	
Pair	Scale	Native	WWT	Native	WWT
En–Ar	360M	0.9%	12.6%	19.53	18.74
En–Ar	7B	5.9%	12.5%	8.45	8.39
En–Ru	360M	1.9%	23.3%	18.99	18.36

Table 2: Merging token spaces by applying WWT to the non-English language raises CL_{eq} and improves English perplexity.

Mapping Arabic to WWT-Ar yields substantial gains in knowledge generalization, raising CL_{eq} to 12.6%, a $14\times$ improvement over the English–Arabic baseline at 360M-parameter scale, and substantially higher than any intervention we test in that disjointly tokenized setting (Fig. 1). Substantial gains appear across our tested settings: at 7B-parameter scale, WWT boosts CL_{eq} from 5.9% to 12.5%, in the English–Russian setting CL_{eq} rises from 1.9% to 23.3% (Tab. 2). We also experiment with partial token space overlap but find it to be ineffective (App.J).

Strikingly, even while training on substantially fewer Arabic documents, the WWT-Ar model's gain in knowledge generalization does not come at the expense of overall English language modeling, with perplexity improving from 19.53 to 18.74. The unified token space allows the model to extract more value from the cross-lingual data it does process, actively driving better English language modeling. Improvements appear across the tested settings (Tab. 2). The knowledge generalization gain is asymmetric, with $CL_{eq_{\text{WWT-Ar}\to\text{English}}} = 17.9\%$ while $CL_{eq_{\text{English}\to\text{WWT-Ar}}} = 7.3\%$. Nonetheless, Arabic language modeling remains surprisingly strong (App. K).

The same asymmetry is present across our experimented scales and language-pairs. We attribute this to the one-directional nature of WWT: WWT-Ar's vocabulary is entirely contained within English tokens, whereas much of English vocabulary never arises as a translation target and is therefore not part of WWT-Ar's vocabulary.

Crucially, WWT-Ar is equivalent to Arabic, therefore, the gain cannot be attributed to WWT-Ar being linguistically closer to English, because in every aspect except tokenization it is not. We conclude that while linguistic differences might dampen cross-lingual knowledge generalization in LLMs, disjoint token spaces are what gate it.

WWT produces substantial knowledge generalization between natural, vocabulary-disjoint languages. It does so without parallel corpora, auxiliary losses, or architectural changes: the intervention is a data-level remapping, applied and inverted in real time. Two limitations follow from fully merging the token spaces: the generation language must be fixed in advance, and code-switched output is unsupported. We address these next.

6.3 Restoring language identity — Soft Mapping

WWT-Ar tokens are indistinguishable from their English counterparts. This is what enables knowledge sharing, but it also means the model has no mechanism for identifying which language it is producing, resulting in the above-mentioned limitations.

Soft mapping relaxes the constraint. Rather than merging tokens, WWT-Ar retains distinct parallel to English token identities and shares a fraction p of its embedding dimensions with the corresponding English tokens, leaving the remainder language-specific (Appendix G).

Replacing WWT-Ar with a high p soft mapping version has a small cost on knowledge generalization with English, consistent with our earlier results on soft-mapping 5.3. CL_{eq} score increases with p and approaches the fully mapped WWT-Ar, reaching $CL_{eq} = 9.60\%$ at $p = 90\%$ and $CL_{eq} = 10.47\%$ at $p = 99\%$ (76% and 83% of the 12.6% obtained under full mapping). A small number of

reserved dimensions therefore preserves most of the knowledge generalization while restoring a native mechanism for language identity, which can even be extended to a multilingual setting, and resolving the limitations of WWT-Ar.

6.4 Semantics matter for token space merging

We test whether the key to our previously introduced setup, mapping Arabic to WWT-Ar, facilitating knowledge generalization lies in token sharing alone, or in the mapping being semantically faithful — each Arabic word mapped to an English word with the same meaning. We evaluate three controls that share tokens without sharing meaning. First, we randomly shuffle the WWT-Ar dictionary, preserving the shared token inventory while destroying the word-level correspondence: CL_{eq} collapses from 12.6% to 3.0% . Second, we map Arabic into English script by transliteration alone. This control is notable because character-level romanization is itself a proposed method for unlocking multilingual capabilities (Moosa et al., 2023; J et al., 2024; Saji et al., 2025): it shares a token inventory without sharing token semantics, and we find that under it CL_{eq} drops to -0.3%. Third, returning to the confound-free $English_1$–$English_2$ setting, we let both copies share the same token space but permute token identities such that the same token index now corresponds to two different subwords: CL_{eq} is -2.6% — same as fully disjoint tokens. Implementation and details for these setups are in App. L

Resolving disjoint tokenization is thus not a matter of token superposition: what knowledge generalization requires is not a shared token inventory but shared *meanings* of tokens. Prior findings that transliteration helps mainly between related languages (J et al., 2024; Jayakumar et al., 2026) are what our mechanism predicts: related languages supply semantic matches through cognates, distant ones do not, making our semantic WWT mapping necessary.

7 Related Work

The Cross-Lingual Knowledge Gap. While LLMs easily acquire linguistic competence from modest data budgets (Charpentier et al., 2025), they struggle to transfer mathematical reasoning, coding, reading comprehension, and other learned skills outside their primary training language (Shi et al., 2023; Peng et al., 2024; Bandarkar et al., 2024; Hu et al., 2020; Cheng et al., 2026). This generalization failure has been shown to extend into the knowledge domain (Goldman et al., 2025; Chang et al., 2026; Guo et al., 2026), and widespread cross-lingual factual inconsistencies have been documented (Qi et al., 2023; Ifergan et al., 2025). Relatedly, Calderon et al. (2026) distinguish missing knowledge from encoded but inaccessible knowledge, finding recall rather than encoding to be the bottleneck. Yet given no definitive answer on where these knowledge gaps originate within the training process, most attempts to mitigate them intervene post-training via prompting and pre-translation (Huang et al., 2023; Wang et al., 2025b; Mondshine et al., 2025; Diskind et al., 2026), knowledge editing (Beniwal et al., 2024), activation manipulation (Wang et al., 2025c;a; Manev, 2026), fine-tuning (Liu et al., 2026; Chua et al., 2025) or reinforcment learning (von Rad et al., 2026b). We instead investigate knowledge generalization in pretraining, where multilingual limitations have been identified before, but in aggregated terms such as model capacity and perplexity (Chang et al., 2024b; Howe & Wingate, 2026). Liu et al. (2025) performed an observational study of multilingual knowledge acquisition on pretraining checkpoints, finding native frequency is the primary driver and characterizing cross-lingual transfer as a second-order effect dependent on script, but their confounded observational setting limits further isolation and understanding of it. Li et al. (2024b) explored cross-lingual failures in pretraining, including knowledge gaps which they find narrow with scale, concluding that alignment emerges spontaneously in larger models. Our controlled measurements challenge this: scaling from 360M to 7B attenuates compartmentalization but leaves the large majority of cross-lingual value unrecovered, and in the confound-free $English_1$–$English_2$ setting scale provides no improvement at all (§5.2). We therefore argue that compartmentalization is established in pretraining rather than dissolved by scale; we attribute this tension to our metric normalizing overall learning ability rather than reporting absolute accuracy (§2.2).

Tokenization as a Generalization Barrier. Typological distance and language family influence transfer (Lauscher et al., 2020), but script match is the primary predictor of whether parametric knowledge successfully crosses language boundaries (Ifergan et al., 2025; Bandarkar et al., 2026). Script reaches the model through tokenization, in two ways. Through fragmentation: non-Latin scripts suffer from over-segmentation (Rust et al., 2021; Petrov et al., 2023), and vocabulary overlap:

Kallini et al. (2025) find that overlap, and especially natural semantic matches, aids generalization. Following on this, existing methods transliterate or merge subwords (Moosa et al., 2023; Zhang et al., 2025; Patil et al., 2022; J et al., 2024; Liu et al., 2024), but their success is limited to closely related languages. Countering both, Conneau et al. (2020b); Artetxe et al. (2020), and Feng et al. (2025) show cross-lingual structures emerge without any shared subwords. We reconcile this: structural alignment does emerge without shared tokens, but knowledge does not follow it (§5.2). Our work corroborates findings on the significance of script by establishing shared token space as the mechanism through which it acts and extends this by effectively applying this principle to distant language pairs.

Pretraining Causal Analysis Tools. Analyzing isolated aspects of multilingual pretraining is confounded by lexical and tokenization-related factors. Prior work introduces synthetic duplicated vocabularies as a simplified test-bed for multilingual analysis (K et al., 2020; Dufter & Schütze, 2020; Schäfer et al., 2024; Howe & Wingate, 2026). A separate line of work studies the dynamics of factual acquisition; this requires knowing exact exposure counts, which is intractable in natural corpora and is therefore estimated (Kandpal et al., 2023; Elazar et al., 2024). A more controlled yet synthetic approach addresses this by tracking acquisition per exposure to fictive facts injected during training (Allen-Zhu & Li, 2024; Chang et al., 2024a). We unify these approaches: we use a cloned-vocabulary testbed, paired with fictive fact injection at known exposure rates, allowing us to measure the effect of isolated interventions and mechanisms on cross-lingual knowledge transfer.

8 Conclusions

Our results converge on a single claim: cross-lingual knowledge compartmentalization originates in pretraining and is in substantial part a tokenization phenomenon. It appears even between languages that differ in nothing but their token inventory, and it is substantially reduced by an intervention that changes nothing but the tokens.

This origin matters for how compartmentalization should be addressed. Most mitigations act after pretraining, restoring access to knowledge the model already holds. They cannot recover what compartmentalization has already cost: a model that learns the same fact separately in each language has spent capacity and compute storing it twice, and output consistency does not undo that duplication.

Tokenization is a substantial part of the story, but not all of it. While token space merging provides significant gains in knowledge generalization, it is far from resolving compartmentalization completely (§5.3). The residual gap plausibly reflects imperfections of our mapping — transliterated out-of-dictionary words and naive conflict-resolving — and the linguistic differences WWT leaves untouched by construction. We leave understanding these factors, possibly utilizing our controlled environment, and improving upon them for future work. We also provide an extended discussion on additional limitations of our work, including token fertility overhead and corpus artifacts, in App. A.

If disjoint token spaces are the barrier, a natural question is whether tokenizer-free architectures dissolve it by construction (Xue et al., 2022; Hwang et al., 2025). Our results do not settle this: a shared byte-level interface removes disjoint token identities, but whether knowledge then generalizes across scripts remains untested, and CL_{eq} offers a direct way to test it.

Ultimately, cross-lingual compartmentalization represents a fundamental flaw in LLM pretraining. Our embedding analysis (§5.2) shows that models successfully learn identical underlying distributions independently, yet completely fail to bridge them — structural similarity is not sufficient for knowledge generalization. This failure has implications beyond text, directly challenging the assumption that multimodal systems will naturally generalize across parallel representations of the same concepts (Huh et al., 2024). Bridging disjoint interfaces, like languages, is a critical step toward disentangling underlying knowledge from its surface form, and moving toward models that learn unified representations of intelligence rather than isolated islands of modality.

AI use statement

In this work, we used generative AI tools for the generation of fictive knowledge datasets as described in §2.1, and for word translation from Russian and Arabic to English as described (App. H). We also made use of generative AI code-assist tools in our development process and used LLMs for feedback during the writing of this paper. We have not used generative AI tools for developing our

main mechanistic claim across this paper, or to support result analysis. We make no mathematical claims, and so tasks related to this do not apply to this work. Additionally, we used generative AI tools to identify relevant literature and summarize or analyse existing literature. We have reviewed all AI-assisted work. We reviewed AI-generated code, verified the contents and relevance of related literature identified by AI, and conducted manual literature reviews. We also verified a sample of the generated fictive knowledge data manually — seed facts, paraphrasings, and evaluations. We take responsibility for the final content of this work, including data, text, and code produced with the aid of generative AI.

Ethics Statement

Our research investigates the fundamental mechanisms of multilingual knowledge transfer to facilitate more equitable and capable AI systems across diverse language communities. The pretraining data used in our experiments consists of standard, publicly available corpora, including FineWeb-edu and FineWeb2-HQ, as well as our Fictional Knowledge Datasets, which are comprised entirely of synthetically generated entities and facts, including human entities that collide with real-human names. We note that these fictive facts are intended only as a probing mechanism, and we do not encourage including them in production-ready pretraining corpora. We also recognize the environmental footprint associated with large-scale pretraining. The majority of our mechanistic investigations and ablation studies were conducted efficiently at the 360M-parameter scale, reserving the compute-intensive 7B-parameter runs for necessary validation only.

Reproducibility Statement

To ensure the full reproducibility of our results, we provide comprehensive documentation and open-source resources. The code for all our experiments, including the synthetically generated fictive knowledge datasets and the evaluation pipeline, are publicly accessible...in our code repository[1]. App. C describes the Fictional Knowledge Dataset (FKD) full generation pipeline in detail, and our published code includes the needed resources to generate it. App. H provides the necessary guidelines on reproducing our WWT operation and can be used with our published code to adapt it for other untested languages. Full technical specifications for our pretraining experiments, including architectural configurations, learning rate schedules, and AdamW optimization hyperparameters, are detailed in App. B. All pretraining was executed using the open-source TorchTitan framework.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=5x788rqbcj.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=FxNNiUgtfa.

Sultan Alrashed, Dmitrii Khizbullin, and David R. Pugh. Fineweb-edu-ar: Machine-translated corpus to support arabic small language models, 2024. URL https://arxiv.org/abs/2411.06402.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.

4623–4637, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Lucas Bandarkar, Alan Ansell, and Trevor Cohn. Large reasoning models struggle to transfer parametric knowledge across scripts, 2026. URL https://arxiv.org/abs/2603.17070.

Himanshu Beniwal, Kowsik D, and Mayank Singh. Cross-lingual editing in multilingual language models. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 2078–2128, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.140. URL https://aclanthology.org/2024.findings-eacl.140/.

Nitay Calderon, Eyal Ben-David, Zorik Gekhman, Eran Ofek, and Gal Yona. Empty shelves or lost keys? recall is the bottleneck for parametric factuality, 2026. URL https://arxiv.org/abs/2602.14080.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/6fdf57c71bc1f1ee29014b8dc52e723f-Abstract-Conference.html.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.236. URL https://aclanthology.org/2024.emnlp-main.236/.

Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, Aditi Gupta, Adril Putra Merin, Adwoa Bremang, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akriti Kuri, Akshay Ramesh, Aleksei Dorkin, Alfred Malengo Kondoro, Alham Fikri Aji, Ali Eren Çetintaş, Allan Hanbury, Alou Dembele, Alp Niksarli, Álvaro Arroyo, Amin Bajand, Amol Khanna, Ana Chkhaidze, Ana Carolina Condez, Anamaria-Roberta Hartl, Andiswa Mkhonto, Andrew Hoblitzell, Andrew Tran, Angelos Poulis, Anirban Majumder, Anjali Chaudhary, Anna Vacalopoulou, Annette Kuuipolani Kanahele Wong, Annika Simonsen, Anton Kovalev, Anupam Nayak, Ashvanth S, Ayodeji Lana, Ayu Purwarianti, Bashar Alhafni, Benedict Busole, Bernard Ghanem, Bharti Nathani, Biljana Stojanovska Đurić, Blessing Ogundipe, Bolaotan Agbonile, Bragi Bergsson, Bruce Torres Fischer, Burak Tutar, Burcu Çınar, Cade Kane, Can Udomcharoenchaikit, Chadi Helwe, Chaithra Reddy Nerella, Chen Cecilia Liu, Chiamaka Nwokolo, Christopher Homan, Clément Sampebgo, Cristina España-Bonet, Cynthia Amol, Daeyoep Lee, Dan Saattrup Smart, Dana Arad, Daniil Dzenhaliou, Dasol Choi, David Liu, David Semedo, David Anugraha, Deborah Popoola, Deividas Mataciunas, Delphine Nyaboke, Dennis Owusu, Dhyuthy Krishna Kumar, Diogo Tavares, Diogo Glória-Silva, Divyanshu Goyal, DongGeon Lee, E. Kelly Buchanan, Ebele Nwamaka Anajemba, Egonu Ngozi Grace, Elena Mickel, Elias Herranen, Eliza Acharya, Eman Nisar, Emile Anand, Emmanuel Habumuremyi, Emuobonuvie Maria Ajiboye, Eryawan Presma Yulianrifat, Esther Adenuga, Ewa Rudnicka, Faith Itiola, Faran Taimoor Butt, Fareeha Fayyaz Sheikh, Fathima Thekkekara, Fatima Haouari, Faustin Nsengiyumva, Fenal Ashokbhai Ilasariya, Filbert Aurelian Tjiaranata, Firas Laakom, Francesca Grasso, Francesco Periti, Francesco Orabona,

Gbenga Kayode Solomon, Genta Indra Winata, Gia Nghia Ngo, Gloria Udhedhe-oze, Gonçalo Vinagre, Gopi Naga Sai Ram Challagolla, Gorka Urbizu-Garmendia, Gouthami Vadithya, Guijin Son, Gulnaz Abdykadyrova, Gyan Swaroop Mohapatra, Hafeez Ullah, Hafsteinn Einarsson, Hai Hu, Hamidreza Saffari, Hamza Zaidi, Haopeng Zhang, Harethah Abu Shairah, Harry Vuong, Hele-Andra Kuulmets, Hitesh Laxmichand Patel, Houda Bouamor, Hwanjo Yu, Iben Nyholm Debess, İbrahim Ethem Deveci, Ikhlasul Akmal Hanif, Ikhyun Cho, Inês Vieira, Inês Calvo, Isaac Manzi, Ismael Illa Salifou, Ismail Daud, Ismail Yusuf, Itay Itzhak, Ivan Zhelyazkov, Ivan Belashkin, Ivan Spada, Jacob Brinton, Jafar Isbarov, Jaka Čibej, Jan Kocoń, Jan Cuhel, Jauza Krito, Jebish Purbey, Jennifer Za, Jennifer Mickel, Jenny Kunz, Jessica Ratovondranto, Jeyarajalingam Varsha, Jihae Jeong, Jimena Tena Dávalos, Jinu Lee, João Magalhães, John Seon Keun Yi, Jongin Kim, Joseph Chataignon, Joseph Marvin Imperial, Jubeerathan Thevakumar, Judith Land, Julia Alekseenko, Junchen Jiang, Jungwhan Kim, Kairit Sirts, Kamesh R, Kamesh V, Kanda Tshinu, Kätriin Kukk, Kaustubh Ponkshe, Kavsar Huseynova, Ke He, Kenneth Enevoldsen, Kent Joshua Alvarez, Kerem Zaman, Khalil Mrini, Kian Kyars, Komal Gour, Krishnakumar Lainitha, Krister Kruusmaa, Kunal Mukherjee, Kusum Chouhan, Laura Castro, Laura M. Porrino-Moscoso, Lenny Sivi Za Nzambi, Leshem Choshen, Levent Sencan, Lilja Øvrelid, Lisa Alazraki, Loretta Oma Jones, Lovina Ehimen-Ugbede, Luheerathan Thevakumar, Luxshan Thavarasa, Mahnoor Malik, Mamadou K. Keita, Mansi Jangid, Marco De Santis, Marcos Garcia, Marek Šuppa, Mariam D'Ciofalo, Marii Ojastu, Marium Attaullah, Maryam Sikander, Mausami Narayan, Maximos Skandalis, Mehak Mehak, Mehmet İlteriş Bozkurt, Melaku Bayu, Menan Velayuthan, Mhasilenuo Vizo, Michael Leventhal, Michał Marcińczuk, Mina Almasi, Mirna Potočnjak, Mithil Bangera, Mohammadamin Shafiei, Mohiba Ansari, Mridul Sharma, Mrityunjaya Indoria, Mughees Ur Rehman, Muhammad Ravi Shulthan Habibi, Murat Kolić, Murat Barkın Kınay, Nada Galant, Naina Singh Rathore, Naphat Permpredanun, Narada Maugin, Nathalie Norman, Nicholas Kluge Corrêa, Nikola Ljubešić, Nirmal Thomas, Nisansa de Silva, Nisheeth Joshi, Nitish Ponkshe, Nizar Habash, Nneoma Udeze, Noel Thomas, Noémi Ligeti-Nagy, Nouhoum Coulibaly, Odunayo Ogundepo, Odunayo Kareemat Buliaminu, Oghojafor Godswill Fejiro, Okechukwu God'spraise, Olanrewaju Samuel, Olaoye Deborah Oluwaseun, Olasoji Akindejoye, Olga Snissarenko, Onyinye Anulika Chiemezie, Orkun Kınay, Osman Tursun, Oyelade Oluwafemi Joshua, Oyesanmi Fiyinfoluwa, Pablo Rodríguez, Pablo Gamallo, Palak Arora, Pedro Valente, Peter Rupnik, Philip Oghenesuowho Ekiugbo, Prakhar Agarwal, Pramit Sahoo, Prokopis Prokopidis, Pua Niau-Puhipau, Quadri Yahya, Rachele Mignone, Raghav Singhal, Rahul Raja, Ram Mohan Rao Kadiyala, Raphael Merx, Rasmus Larsen, Ratnavel Rajalakshmi, Rishav Ghosh, Romina Oji, Ron Kekeha Solis, Rui Guerra, Rushikesh Zawar, Sa'ad Nasir Bashir, Saeed Alzaabi, Sahil Sandeep, Sai Pavan Batchu, Sai Sandeep Kantareddy, Saleha Muzammil, Salsabila Zahirah Pranida, Sam Buchanan, Samuel Rutunda, Sander Land, Sarah Sulollari, Sardar Ali, Saroj Sapkota, Sarveswaran Kengatharaiyer, Saulius Tautvaisas, Sayambhu Sen, Sayantani Banerjee, Sebastien Diarra, Segun Afolayan, Senthilnathan M, Sewoong Lee, Shaan Shah, Shankar Venkitachalam, Sharifa Djurabaeva, Sharon Ibejih, Shivanya Shomir Dutta, Siddhant Gupta, Silvia Paniagua Suárez, Sina Ahmadi, Sivasuthan Sukumar, Siyuan Song, Snegha A, Sokratis Sofianopoulos, Sona Elza Simon, Sonja Benčina, Sophie Gvasalia, Sphurti More, Spyros Dragazis, Stefan Milosavljević, Stephan P. Kaufhold, Suba S, Sultan Alrashed, Surangika Ranathunga, Taiga Someya, Taja Kuzman Pungeršek, Tal Haklay, Tasi'u Jibril, Tatsuya Aoyama, Tea Abashidze, Terenz Jomar Dela Cruz, Terra Blevins, Themistoklis Nikas, Theresa Idoko, Thu Mai Do, Tilek Chubakov, Tina Munda, Tobiloba Owoeye, Tommaso Gargiani, Uma Rathore, Uni Johannesen, Uwuma Ugwu, Vallerie Alexandra Putra, Vanya Bannihatti Kumar, Varvara Arzt, Vasily Konovalov, Vasudevan Nedumpozhimana, Viktoria Ondrejova, Viktoryia Horbik, Vishnu Vardhan Reddy Kummitha, Vuk Dinić, Walelign Sewunetie, Winston Wu, Xiaojing Zhao, Yacouba Diarra, Yaniv Nikankin, Yash Mathur, Yash Bagla, Yeshil Bangera, Yixi Chen, Yiyuan Li, Yolanda Xavier, Yonatan Belinkov, Zaid Alyafeai, Zhargal Batozargalova, Zhengyang Shan, Zhi Rui Tam, Zilu Tang, Zuzana Nadova, Baber Abbasi, Stella Biderman, David Stap, Duygu Ataman, Fabian Schmidt, Hila Gonen, Jiayi Wang, and David Ifeoluwa Adelani. Global piqa: Evaluating commonsense reasoning across 100+ languages and cultures, 2026. URL https://arxiv.org/abs/2510.24081.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen,

Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams (eds.), *Proceedings of the First BabyLM Workshop*, pp. 399–420, Suzhou, China, 2025. Association for Computational Linguistics. ISBN TODO. doi: 10.18653/v1/2025.babylm-main.28. URL https://aclanthology.org/2025.babylm-main.28/.

Bobby Cheng, Adam Gaber, Zhengyuan Liu, Catherine Arnett, Omer Goldman, Cheston Tan, and Leshem Choshen. Skill issue: Are skills language-invariant in llms?, 2026.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL https://aclanthology.org/2021.naacl-main.280/.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. Crosslingual capabilities and knowledge barriers in multilingual large language models, 2025. URL https://arxiv.org/abs/2406.16135.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL https://aclanthology.org/2020.acl-main.536/.

Elisha Diskind, Itamar Trainin, Uri Shaham, Leshem Choshen, Idan Szpektor, and Omri Abend. Cross-lingual exploration for parametric knowledge, 2026. URL https://arxiv.org/abs/2606.24579.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. Pretraining language models using translationese. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5843–5862, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.334. URL https://aclanthology.org/2024.emnlp-main.334/.

Philipp Dufter and Hinrich Schütze. Identifying elements essential for BERT's multilinguality. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.358. URL https://aclanthology.org/2020.emnlp-main.358/.

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. What's in my big data? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=RvfPnOkPV4.

Zihao Feng, Hailong Cao, Wang Xu, and Tiejun Zhao. Word-level cross-lingual structure in large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2026–2037, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.138/.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 2024. URL https://zenodo.org/records/12608602.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer, 2025. URL https://arxiv.org/abs/2502.21228.

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, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer,

Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=uz0gAAYydl.

Pei-Fu Guo, Yun-Da Tsai, Chun-Chia Hsu, Kai-Xin Chen, Ya An Tsai, Kai-Wei Chang, Nanyun Peng, Mi-Yen Yeh, and Shou-De Lin. LiveCLKTBench: Towards reliable evaluation of cross-lingual knowledge transfer in multilingual LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15193–15208, San Diego, California, United States, 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.694. URL https://aclanthology.org/2026.acl-long.694/.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza

Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6997–7013, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482/.

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

Thomas Vincent Howe and David Wingate. Language models struggle with compartmentalization, 2026. URL https://arxiv.org/abs/2605.19284.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hu20b.html.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12365–12394, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.826. URL https://aclanthology.org/2023.findings-emnlp.826/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Sukjun Hwang, Brandon Wang, and Albert Gu. Dynamic chunking for end-to-end hierarchical sequence modeling, 2025. URL https://arxiv.org/abs/2507.07955.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4630–4644, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.475/.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15593–15615, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.833. URL https://aclanthology.org/2024.acl-long.833/.

Thanmay Jayakumar, Deepon Halder, and Raj Dabre. Scripts through time: A survey of the evolving role of transliteration in nlp, 2026. URL https://arxiv.org/abs/2604.18722.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJeT3yrtDr.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. False Friends are not foes: Investigating vocabulary overlap in multilingual language models. In Christos Christodoulopoulos,

Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21138–21154, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1153. URL https://aclanthology.org/2025.findings-emnlp.1153/.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4483–4499, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.363. URL https://aclanthology.org/2020.emnlp-main.363/.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/9a16935bf54c4af233e25d998b7f4a2c-Abstract-Conference.html.

Jiahuan Li, Shujian Huang, Aarron Ching, Xinyu Dai, and Jiajun Chen. PreAlign: Boosting cross-lingual transfer by early establishment of multilingual alignment. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10246–10257, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.572. URL https://aclanthology.org/2024.emnlp-main.572/.

Wanchao Liang, Tianyu Liu, Less Wright, Will Constable, Andrew Gu, Chien-Chin Huang, Iris Zhang, Wei Feng, Howard Huang, Junjie Wang, Sanket Purandare, Gokul Nadathur, and Stratos Idreos. Torchtitan: One-stop pytorch native solution for production ready LLM pretraining. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=SFN6Wm7YBI.

Tianyu Liu, Jirui Qi, Mrinmaya Sachan, Ryan Cotterell, Raquel Fernández, and Arianna Bisazza. Post-training language models for crosslingual consistency, 2026. URL https://arxiv.org/abs/2603.04678.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2476–2499, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.136. URL https://aclanthology.org/2024.acl-long.136/.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schütze. Tracing multilingual factual knowledge acquisition in pretraining, 2025. URL https://arxiv.org/abs/2505.14824.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Alexander Manev. Inference-time steering for cross-lingual factual consistency in llms, 2026. URL https://arxiv.org/abs/2607.19243.

Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. Enhancing multilingual llm pretraining with model-based data selection. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2109. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5aee125f052c90e326dcf6f380df94f6-Paper-Datasets_and_Benchmarks_Track.pdf.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1331–1354, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.73. URL https://aclanthology.org/2025.findings-naacl.73/.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. Does transliteration help multilingual language modeling? In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 670–685, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.50. URL https://aclanthology.org/2023.findings-eacl.50/.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 219–233, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.18. URL https://aclanthology.org/2022.acl-long.18/.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin A. Raffel, Leandro von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Qiwei Peng, Yekun Chai, and Xuhong Li. HumanEval-XL: A multilingual code generation benchmark for cross-lingual natural language generalization. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8383–8394, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.735/.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/74bb24dca8334adce292883b4b651eda-Abstract-Conference.html.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5877–5891, Toronto,

Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.323. URL https://aclanthology.org/2023.acl-long.323/.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10650–10666, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.658. URL https://aclanthology.org/2023.emnlp-main.658/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. RomanLens: The role of latent Romanization in multilinguality in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26410–26429, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1354. URL https://aclanthology.org/2025.findings-acl.1354/.

Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. The role of language imbalance in cross-lingual generalisation: Insights from cloned language experiments, 2024. URL https://arxiv.org/abs/2404.07982.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=fR3wGCk-IXp.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Jonathan von Rad, Louis Arts, George Burgess, Eleftheria Kolokytha, Harry O'Donnell, Ektor Oikonomidis Doumpas, Eduardo Sanchez, Yao Lu, and Pontus Stenetorp. Improving cross-lingual factual recall via consistency-driven reinforcement learning, 2026a. URL https://arxiv.org/abs/2606.06586.

Jonathan von Rad, Louis Arts, George Burgess, Eleftheria Kolokytha, Harry O'Donnell, Ektor Oikonomidis Doumpas, Eduardo Sanchez, Yao Lu, and Pontus Stenetorp. Polyfact: Comparing consistency-driven post-training methods for cross-lingual factual recall, 2026b. URL https://arxiv.org/abs/2606.06586.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5075–5094, Vienna, Austria, 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.253. URL https://aclanthology.org/2025.acl-long.253/.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. Multilingual prompting for improving LLM generation diversity. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6367–6389, Suzhou, China, 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.324. URL https://aclanthology.org/2025.emnlp-main.324/.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5418–5433, Vienna, Austria, 2025c. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.270. URL https://aclanthology.org/2025.acl-long.270/.

Shijie Wu and Mark Dredze. Do explicit alignments robustly improve multilingual encoders? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4471–4482, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.362. URL https://aclanthology.org/2020.emnlp-main.362/.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17/.

Linda Zeng, Steven Y. Feng, and Michael C. Frank. Bringing up a bilingual babylm: Investigating multilingual language acquisition using small-scale models, 2026. URL https://arxiv.org/abs/2603.29552.

Crystina Zhang, Jing Lu, Vinh Q. Tran, Tal Schuster, Donald Metzler, and Jimmy Lin. Tomato, tomahto, tomate: Do multilingual language models understand based on subword-level semantic concepts? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1821–1837, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.98. URL https://aclanthology.org/2025.findings-naacl.98/.

A Limitations

Corpus Artifacts. Our Arabic pretraining corpus is machine-translated from English. This risks introducing "translationese": structural and syntactic artifacts that make the Arabic text artificially resemble English grammar, potentially inflating the baseline structural alignment (Doshi et al., 2024). We mostly mitigate this by replicating our core compartmentalization and token-space unification findings on native Russian data.

Linear Measurement of Non-Linear Learning. The CL_{eq} score is estimated using a linear approximation. This sits in tension with the log-like shape of the knowledge acquisition curve (App. C.5). While we address this choice in §2.2, it inherently oversimplifies the true dynamics of memorization. Our FKD dataset is limited in complexity — relatively simple entity–attribute facts — and size — 2048 facts. The latter limitation results in measurement noise, discussed extensively in App. E, where we also provide significance measurements of our headline comparisons.

Inference Overhead from Token Fertility. While the WWT mapping operation itself is computationally cheap, it causes a substantial increase in sequence length, increasing the cost of inference. We attribute this token fertility increase primarily to our conflict-resolution strategy (App. H). This overhead could potentially be mitigated by a more refined dictionary acquisition process or a better conflict resolution policy. However, because the primary contribution of this mapping in our scope is mechanistic, we leave the engineering of a fully optimized mapping to future work.

Scalability to Massively Multilingual Settings. We evaluate token space merging in a bilingual context. We do not address the capacity limits of a single token space, leaving open the question of how many languages can be mapped into the English token space and how those would interfere.

B Training Details

All models are pretrained using the TorchTitan framework (Liang et al., 2025). An architectural breakdown and hyperparameters used can be found in Tab. B.1

B.1 Architectures

SmolLM2-360M. A standard dense transformer following the SmolLM2 architectural configuration (Allal et al., 2025).

Llama-3 7B. A scaled-down adaptation of the Llama-3 8B architecture (Grattafiori et al., 2024). We reduce the number of hidden layers from 32 to 28 and the intermediate feed-forward dimension from 14,336 to 13,312, giving approximately 7 billion parameters.

B.2 Tokenizer

We use a BPE tokenizer with a vocabulary of 65,536 tokens, trained on a 50/50 split of our bilingual data. For English_1–English_2 experiments, we double the vocabulary to 131,072 to accommodate two fully disjoint language copies.

B.3 Optimisation

All models are optimised with AdamW (Loshchilov & Hutter, 2019) using $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay 0.1. The learning rate follows a warmup–decay, cosine decay, schedule.

B.4 Scale-Specific Configurations

Compute budgets target Chinchilla-optimal data scaling (Hoffmann et al., 2022). Learning rates are inspired by the original architecture configurations and tuned to our use case.

SmolLM2-360M. 4,600 steps, batch size 768, sequence length 2,048, for approximately 7.2B tokens. Warmup stage is 300 steps, peaking at lr 5×10^{-4} then decaying to 2.5×10^{-5}. Trained on 8 RTX B6000 Pro GPUs.

Llama-3 7B. 133,600 steps, batch size 512, sequence length 2,048, for approximately 140B tokens. Warmup stage is 1,000 steps peaking at lr 3×10^{-4} then decaying to 3×10^{-5}. Models are trained on 128 H100 GPUs across 16 nodes.

B.5 Injection Mechanics

To compute CL_{eq} we split FKD into 16 groups of 128 entities. Each group is assigned an exposure count for language A and language B drawn from the cross-product of $\{0, 20, 100, 1000\}$ with itself. Documents are injected stochastically during pretraining at a per-language probability matching the target exposure count for that fact. We report the realized injection counts and SDs in Tab. B.2 based on a representative run. We find this pattern replicates across runs.

B.6 Two-Stage Mixing Schedule

When language mixing intervention — code-switching — is applied at 360M parameter scale, training is divided into two stages. The first and longer stage (3,600 steps) applies the mixing at the chosen mixing rates; the second, shorter stage applies none, realigning the model to the natural data distribution and preventing excessive code-switching in generation.

C Fictive Knowledge Dataset (FKD)

This appendix details the construction of the fictive knowledge dataset — FKD— we inject into pretraining to measure cross-lingual knowledge generalization (§2.1). Because the facts are fictive,

	SmolLM2-360M	Llama-3 7B
Architecture		
Layers	32	28
Hidden dimension	960	4,096
FFN intermediate dim.	2,048	13,312
Attention heads	15	32
KV heads (GQA)	5	8
Vocabulary	65,536	65,536
Tied input/output emb.	Yes	No
Optimisation		
Optimiser	AdamW	AdamW
β_1, β_2	0.9, 0.95	0.9, 0.95
Weight decay	0.1	0.1
Peak LR	5×10^{-4}	3×10^{-4}
Final LR	2.5×10^{-5}	3×10^{-5}
Schedule	Cosine	Cosine
Warmup steps	300	1,000
Data		
Steps	4,600	133,600
Batch size (seqs)	768	512
Sequence length	2,048	2,048
Tokens per batch	$\sim$1.57M	$\sim$1.05M
Total tokens	$\approx$7.2B	$\approx$140B
Hardware		
GPUs	8$\times$RTX B6000 Pro	128$\times$H100 (16 nodes)
GPU hours	24	12,000–15,000

Table B.1: Pretraining configurations at both scales. Every CL_{eq} result in the paper corresponds to one full pretraining run under one of these two configurations. The $English_1$–$English_2$ runs instead use a doubled vocabulary of 131,072 tokens, to accommodate two disjoint token spaces (see *Tokenizer*, above).

Target	Mean realised	SD	Relative SD
0	0	0	—
20	20.6	5	24.3%
100	101.2	11	10.9%
1000	1011.8	32	3.2%

Table B.2: Realised exposure counts against their targets, measured on a representative 360M English–Arabic run. Deviations are consistent with the stochastic injection procedure; relative deviation is largest in the low-exposure cells.

they are guaranteed to be absent from the underlying corpus, so all recall is attributable to our injections and the number of exposures in each language is known exactly. The dataset and generation code are publicly available[1].

FKD covers 3 languages that we evaluate in this paper — English, Arabic and Russian. It is generated in three stages: seed facts, injection documents, and evaluation multi-choice questions (MCQs). Tab. C.1 provides a representative example.

C.1 Stage 1: Seed Facts

A seed fact pairs a named entity with a single attribute (e.g., *"Neltis is a podcast about the business of coffee"*). We manually author 64 fact templates spanning entity types such as companies, podcasts, and tools. Each template is passed to Gemini 3.1 Pro with instructions to produce 32 distinct facts

using artificially generated entity names (with the exception of entity types that are human, where generating some generic name), yielding $64 \times 32 = 2{,}048$ seeds.

Translation and entity names. Seeds are translated into all target languages with entity names transliterated. We use GPT-5.5-mini at a low reasoning setting for this end. This means entity names share no tokens across scripts, avoiding this confounding variable in knowledge generalization (Bandarkar et al., 2026).

Language balance. Generated names inherit the phonotactics of the language they were generated in: prompting in English yields names that read as phonetically plausible in English but not necessarily elsewhere. We verified this by manual inspection. To prevent the resulting bias from favouring a single language, we generate seeds in equal proportion across English and Arabic, the pair on which most of our main experiments focus.

C.2 Stage 2: Injection Documents

Facts in real corpora appear scattered and paraphrased across heterogeneous contexts rather than in a single canonical statement. To reproduce this, we expand each translated seed into many short documents that state or imply the fact in varied natural settings: casual blog posts, news items, advertisements, forum replies, and similar. Expansion uses GPT-5.5-mini at a low reasoning setting, with identical prompts across languages. We generate 200–300 documents per entity per language. The resulting documents' average token length is ∼20 tokens.

This diversity is deliberate; Allen-Zhu & Li (2024) show that knowledge must be sufficiently augmented during pretraining—through paraphrasing and varied presentation—to acquire generalizable knowledge. While their experiments conclude that even 5 augmentations are enough, we choose to target a higher number to better capture natural settings.

C.3 Stage 3: Evaluation Questions

For each fictive fact and language, we generate five four-choice questions testing knowledge of it. Distractors are generated to be as plausible as the correct choice given no prior knowledge but clearly incorrect given the fact. These questions are generated separately from the injection documents and never appear in the training corpus. We evaluate using the LM Evaluation Harness (§3.2), averaging over the five questions per entity.

C.4 FKD for WWT-Ar

FKD for WWT-Ar introduced in §6.1 is obtained by applying the corresponding WWT mapping to the Arabic documents and evaluations, so that WWT-Ar contains the same facts, documents and evaluations as natural Arabic.

C.5 FKD Validation & Fact-Acquisition Learning Curve

We validate our fictive facts datasets - FKD- and injection pipeline by pretraining 360M-parameter models with monolingual FKD injection. We vary the exposure rates and measure fact acquisition using the accompanying MCQs; results in Fig. C.1.

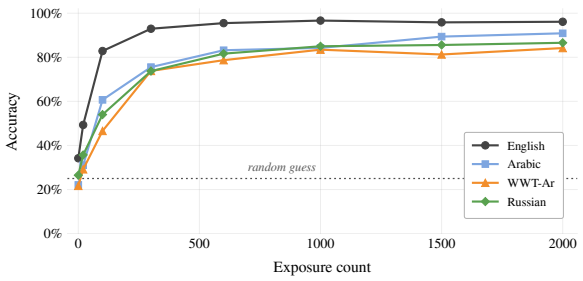

Figure C.1: Monolingual acquisition of fictive facts (FKD). The graph depicts the performance (y-axis) per number of times a fact was seen (x-axis) per input language (line color).

Our measurements reveal that fact acquisition (measured through MCQ accuracy) exhibits a log-like dependence on exposure count. Accuracy increases rapidly at the lower exposure range (0-300 range) and plateaus around

1000 exposures. Our evaluation injection grid is concentrated in the steep region, where learning efficiency is most measurable (§3.2). While the overall learning-curve trend is consistent across languages, Accuracy values differ, possibly reflecting a bias in our generation process. Nonetheless, our metric's definition as relative gain eliminates these differences from our measurements (§2.2).

Our knowledge acquisition measurements are consistent with prior work on characterizing learning dynamics under controlled settings and extend them into the natural regime. Allen-Zhu & Li (2025) measure knowledge acquisition in synthetic setups, with many repetitions and a predefined document structure, and report that 1000 fact exposures reach near-perfect memorization. Extending this analysis to natural settings is challenging as knowledge exposure is intractable in enormous pretraining corpora. Thus, natural-corpus analyses evaluate fact acquisition and estimate exposure count based on entity frequency, reporting substantially slower acquisition (Kandpal et al., 2023), possibly due to overcounting when the entity is mentioned but the fact is not. Our controlled setup using FKD provides precise measurement in a natural setting, confirming the synthetically measured dynamics.

C.6 Examples

FKD	
Seed	Arqous is a spice extracted from celery stalks.
English Document	A little Arqous, the celery-stalk derived spice, goes a long way in soups and stews.
English MCQ	Arqous is a seasoning extracted from (a) celery stalks (b) carrot stalks (c) parsley stalks (d) spinach stalks

Table C.1: Representative example from FKD. All seeds, documents, and MCQs in the three supported languages appear in the released data.

D Human Fictional Knowledge Dataset (HFKD)

FKD mostly uses artificially generated entity names, placing its entities names in the low-frequency token regime. Prior work indicates that cross-lingual bridging mechanisms behave differently for high- and low-frequency tokens (Schäfer et al., 2024), so we construct a second dataset, Human Fictional Knowledge Dataset (HFKD), using human entities with common first and last names—tokens that appear frequently in any pretraining corpus. All downstream processing (document expansion, evaluation questions, injection) is identical to FKD.

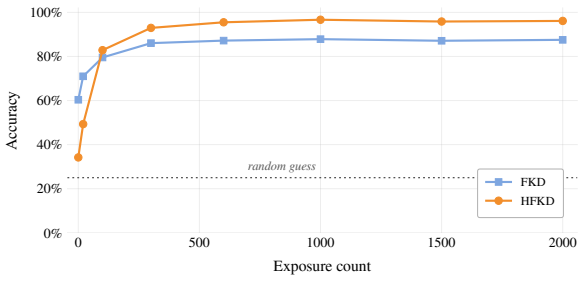

Figure D.1: Monolingual English acquisition of HFKD and FKD. The graph depicts the performance (y-axis) per number of times a fact was seen (x-axis) per fictive knowledge dataset (line color).

D.1 Seed Generation

To achieve diversity for human entities, we generate the fact space compositionally. We define 50 professional domains (cinema, science, etc.); GPT-5.5-mini generates 5 roles per domain (for cinema: actor, director, etc.) and 4 attributes per role (for actor: awards, debut, genre, etc.), giving 1,000 role–attribute templates. Each seed fact samples a first and last name from curated lists of 100 each, pairs them with a template, and prompts the model to fill in a plausible attribute value. We sample 2,048 such combinations.

Setting	FKD	HFKD
English–Arabic		
Baseline	0.9%	1.3%
Language mixing	2.2%	1.7%
Activation alignment	2.5%	2.2%
WWT	12.6%	8.7%
English$_1$–English$_2$		
Basline	−2.6%	−8.3%
Soft tying, $p = 20\%$	8.6%	4.0%
Soft tying, $p = 50\%$	82.7%	57.9%
Soft tying, $p = 90\%$	95.8%	87.8%
Soft tying, $p = 99.9\%$	99.0%	101.6%
Shared Initialization	78.6%	59.4%

Table D.1: CL_{eq} on FKD and HFKD across the paper's main settings on a 360M-parameter scale. Ordering is mostly preserved throughout, and headline comparisons hold;

D.2 Acquisition Characteristics

HFKD is not directly comparable to FKD, Its monolingual acquisition curve occupies a substantially compressed range (Fig. D.1), beginning at 60% and plateauing around 87%, against FKD's 34% to 96%. We suspect this is the result of contamination from real individuals, or generation quality. We therefore interpret HFKD scores with caution and provide them only as an indication that our main results are consistent across frequency regimes.

D.3 Replication Results

Every qualitative conclusion in the paper replicates on HFKD: bilingual pretraining is heavily compartmentalized. the English$_1$–English$_2$ setup shows poor knowledge generalization as well, with embedding tying restoring it monotonically. Finally, WWT substantially improves knowledge generalization between natural vocabulary disjoint languages (Table D.1).

HFKD estimates are substantially noisier, appearing miscalibrated at the metric floor with one of the scores well into the negative region. This follows from its compressed acquisition range: because accuracy spans roughly 30 points rather than 60, the native-exposure slope β_A is correspondingly smaller, and since CL_{eq} is the ratio β_B/β_A, a smaller denominator inflates the variance of the estimate while the entity-based noise in β_B is unchanged. The negative values should therefore be read as absolute compartmentalization against a wider noise floor than FKD's, not as negative transfer.

E Cross-lingual Equivalence Score (CL_{EQ})

This appendix documents how CL_{eq} is estimated, why the linear specification is preferred over a curve-aware one, and how much sampling variance the estimates carry — establishing that the paper's headline comparisons survive that variance.

E.1 Estimation and Fit Diagnostics

We estimate CL_{eq} by estimating native and foreign exposure value to fact acquisition through ordinary least squares over all facts in FKD, regressing language-A accuracy on the injection counts in both languages (§2.2). Native learning is strong across our experiments: β_A is positive and significant throughout, so the denominator of the ratio is bounded away from zero and CL_{eq} estimates do not exhibit the instability characteristic of small-denominator ratio estimators. Precision is nonetheless limited by the numerator, as we discuss below.

Table E.1 reports the fitted coefficients for two representative runs at 360M: the English–Arabic baseline and its WWT-Ar counterpart, both evaluated in the Arabic→English direction. Native learning is comparable across the two ($\beta_A \approx 4.4$–4.9×10^{-4} accuracy gain per exposure, accuracy range is 0–1), so the difference in CL_{eq} is driven almost entirely by the cross-lingual coefficient.

Setting (A–B)	$\text{CL}_{\text{eq}\,B\to A}$	α	β_A	β_B	**95% CI** (β_B)	R^2
English–Arabic	0.64%	0.489	4.90×10^{-4}	3.12×10^{-6}	$[-2.9,\ 3.6] \times 10^{-5}$	0.66
English–WWT-Ar	17.9%	0.517	4.44×10^{-4}	7.96×10^{-5}	$[4.8,\ 11.2] \times 10^{-5}$	0.60

Table E.1: Fitted coefficients of Eq. 1 for two representative 360M runs, in the Arabic→English direction. α is the intercept, β_A and β_B the accuracy gained per native and per cross-lingual exposure. Native learning is comparable across settings; the baseline's cross-lingual coefficient is indistinguishable from zero, while WWT-Ar's is significantly positive. Notice that Eq. 1 fits accuracy between 0 and 1, hence the scale of the diffrent fitted parameters.

In the baseline, β_B is statistically indistinguishable from zero: its 95% confidence interval spans zero (CI across FKD fact samples). Under WWT-Ar, β_B is significantly positive and its interval lies entirely above the baseline's upper bound.

The linear specification captures most but not all of the variation in accuracy (R^2 column in Tab. E.1). The residual misfit is systematic rather than random. This is the expected consequence of fitting a saturating curve linearly over a non-uniform exposure grid — the fit over-predicts at zero exposures and under-predicts at 100 — and it motivates a curve-aware specification.

E.2 Curve-Aware CL_{EQ}

Fact acquisition follows a log-like rather than linear dependence on exposure count (§2.1), pointing towards a misfit of our linear fit estimator. We therefore also implement a curve-aware estimator that models saturation directly:

$$\text{acc}_A(f) = a \log\big(exps_A(f) + \alpha\, exps_B(f) + 1\big) + b, \tag{4}$$

fitted by nonlinear least squares over three free parameters. Here α admits a direct reading as the CL_{eq} quantity: it is the number of native exposures that one foreign exposure is worth, entering the curve on the same footing as $exps_A$. The curve-aware score is therefore $100 \times \alpha$.

The curve-aware specification fits the data substantially better, but its α is not well identified. Because α enters only inside the logarithm alongside $exps_A$, and because a scales and b shifts the resulting curve, the three parameters trade off against one another. This additional flexibility improves fit without guaranteeing that α isolates the quantity of interest.

The linear specification has the opposite character. It fits less well, and its residuals retain systematic curvature as we mentioned above, but β_B multiplies $exps_B$ alone and is therefore identified by construction: no combination of the remaining parameters can absorb variation along the foreign axis. We prefer this trade—a worse fit with an interpretable parameter over a better fit with a confounded one—and report the linear estimator throughout. Table E.2 reports both, while absolute values differ, both specifications provide the same ranking among our main comparisons.

E.3 Sampling Variance and Significance

CL_{eq} is a ratio estimator computed over a finite set of entities, and its sampling variance is substantial. Native learning is robust in every experiment we report, so the denominator contributes little noise; the variance is carried almost entirely by the numerator, which in most conditions estimates a very weak effect. The suspect candidates for the noise sources are per-fact variation in learnability, in the quality of its paraphrases, and in the difficulty of its evaluation questions.

Estimated independently, the resulting intervals are wide. At 360M the English–Arabic baseline gives $\text{CL}_{\text{eq}} = 0.9\%$ with a 95% CI of $[-5.5, 7.4]$, while WWT-Ar gives 12.6% with $[5.4, 20.0]$. These marginal intervals overlap, but they answer the wrong question. Much of the variance is shared: a resampled set of entities that results in an extreme CL_{eq} value for one model effects other models similarly, so entity variance inflates each interval without affecting the comparison between methods.

We therefore resample entities *jointly* across conditions, drawing one bootstrap sample per replicate (within-cell, entity-wise, with replacement across our 16 groups of 128 entities), evaluating all conditions on that same sample, and reporting the distribution of *differences* (Table E.3). Because

	Linear		Curve-aware	
Setting	$\mathrm{CL}_{\mathrm{eq}}$	R^2	$\mathrm{CL}_{\mathrm{eq}}$	R^2
English–Arabic 360M				
Baseline	0.9%	0.77	0.15%	0.88
WWT	12.6%	0.73	0.49%	0.88
English–Arabic 7B				
Baseline	5.9%	0.58	0.09%	0.92
WWT	12.5%	0.58	0.28%	0.90
English$_1$–English$_2$ 360M				
Baseline	−2.6%	0.71	0.09%	0.91
Soft tying, $p = 99.9\%$	99.0%	0.51	87.7%	0.90
Shared initialization	78.6%	0.55	45.1%	0.89
English$_1$–English$_2$ 7B				
Baseline	−1.5%	0.70	0.11%	0.87

Table E.2: $\mathrm{CL}_{\mathrm{eq}}$ under the linear and curve-aware estimators. The curve-aware fit attains higher R^2 throughout, while the ordering of conditions is mostly preserved under both.

every condition sees the same entities in a given replicate, shared entity variance cancels, and what remains is the variance of the comparison itself.

Under this procedure, the comparisons supporting our central claims separate cleanly. WWT-Ar exceeds the English–Arabic baseline by 11.7 percentage points, with a 95% CI of $[7.5, 16.1]$ that excludes zero; even its lower bound is several times the baseline point estimate. The same holds at 7B and for the semantic controls of §6.4.

Two consequences follow for reading the paper's numbers. First, small negative $\mathrm{CL}_{\mathrm{eq}}$ estimates arise from noise in β_B when cross-lingual learning is very weak; they are indistinguishable from zero and should be read as substantial or absolute compartmentalization rather than as negative transfer. Second, differences of a few percentage points between conditions should not be interpreted as meaningful — the comparisons we draw conclusions from are those in Table E.3 whose intervals exclude zero.

Scope of the bootstrap. This procedure captures evaluation-side variance only. It does not capture variance across training runs: each configuration is a single pretraining run, and both the realized injection counts and their placement in the training order differ between runs. Repeating every configuration across seeds is prohibitive at these scales. We instead argue robustness through replication across the axes we can vary: model scale, language pair, and the close to monotone dose-response in the English$_1$–English$_2$ tying (§5.3) and English–Arabic soft-mapping (§6.3) sweeps, where noise alone would not produce a consistent trend across independent runs.

E.4 Target and Realized Exposure

Because injection is stochastic, the realized number of exposures deviates from the target. Deviations are consistent across our experiments, values reported in Tab. B.2. In relative terms the low-exposure cells are the most affected—24.3% at 20 exposures against 3.2% at 1000—and these are precisely the cells in the steepest region of the acquisition curve (App. C.5), where accuracy is most sensitive to exposure count.

We estimate $\mathrm{CL}_{\mathrm{eq}}$ using target exposures. To verify that this does not distort our measurements, we recompute $\mathrm{CL}_{\mathrm{eq}}$ using the realized counts for two representative runs and find the difference negligible (Table E.4).

F Intervention Implementation Details

This appendix provides implementation details and detailed results for the interventions evaluated in §4.2.

Comparison	Δ	95% CI	Significant
English/Arabic, 360M			
WWT-Ar vs. baseline	+11.7	[7.5, 16.1]	*
WWT-Ar vs. Shuffled WWT-Ar	+9.6	[4.5, 14.7]	*
WWT-Ar vs. Transliterated-Ar	+13.0	[8.9, 17.2]	*
English/Arabic, 7B			
WWT-Ar vs. baseline	+6.6	[1.8, 11.4]	*
English/Russian, 360M			
WWT-Ru vs. baseline	21.3%	[16.8, 25.9]	*
WWT-Ru vs. Shuffled WWT-Ru	19.7%	[13.3, 26.1]	*
$English_1 - English_2$, *360M*			
99.9% tying vs. baseline	+101.6	[91.5, 113.5]	*
99.9% tying vs. shared initialization	+20.4	[12.8, 28.6]	*

Table E.3: Paired-bootstrap differences in CL_{eq} for the paper's main comparisons. Each replicate resamples entities once and evaluates all conditions on the same sample, so shared entity-difficulty variance cancels. * marks intervals excluding zero.

Setting	Target	Realized
English–Arabic Baseline	0.9%	1.3%
English–Arabic WWT	12.6%	13.1%

Table E.4: CL_{eq} estimated from target versus realized exposure counts. Using realized counts leaves both the values and their difference essentially unchanged.

F.1 Language Mixing — Code-switching

Inspired by lingual code-switching, pretraining code-switching introduces cross-lingual context by randomly swapping words with their cross-lingual counterparts during pretraining. Because the language of the next token becomes stochastic, the model is pushed toward assigning shared representations across languages. We vary two parameters: the *mixing ratio*—the proportion of documents subjected to code-switching—and the *substitution probability*—the probability of replacing any given word within such a document. We execute the substitution using a variant of our curated dictionary presented in (App. H), which skips the conflict resolution stage to avoid the appended IDs on substituted words.

Data exclusion. We strictly exclude FKD documents from the mixing pool, simulating the realistic case in which a limited aligned dataset is used alongside a much larger body of standard-format data for which we would like to evaluate transfer.

Results. We sweep the mixing ratio over $\{0.6, 0.1, 0.01\}$ and the substitution probability over $\{0.5, 0.1, 0.01\}$ (Table F.1). No setting produces meaningful knowledge generalization: CL_{eq} ranges from -1.2% to 2.2% across the sweep, against a baseline of 0.9% (App. E). There is no clear trend in either parameter — the best and worst cells differ in mixing ratio by a factor of 60, but the ordering is not monotone in either direction, indicating the differences can be mainly attributed to noise. Code-switching in a natural English–Arabic vocabulary-disjoint setting therefore does not appear to induce knowledge sharing at any mixing volume we test.

F.2 Activation Alignment Loss

Activation Alignment loss encourages the model to align inner representations of cross-lingual semantically equivalent units. We extract a global representation of a sequence by concatenating the max and average pooling of its layer-4 activations. We also experiment with multi-depth aggregation of the loss on the SmolLM2-360M architecture, where we average the losses from layers $\{4, 8, 12, 16, 20, 24, 28\}$ (having normalized the losses to account for differences in scales across

	Mixing ratio		
Subst. prob.	**0.6**	**0.1**	**0.01**
0.5	2.2%	0.9%	−1.2%
0.1	1.6%	0.3%	2.0%
0.01	−0.4%	1.5%	0.4%

Table F.1: CL_{eq} under pretraining code-switching in the English–Arabic setting at 360M, across mixing ratio and substitution probability. The native-script baseline is 0.9%.

	Chunk size n (words)			
Loss	**4**	**20**	**100**	**250**
InfoNCE	2.30%	2.22%	2.54%	2.23%
L2	1.89%	1.69%	0.32%	0.47%

Table F.2: CL_{eq} under activation alignment in the English–Arabic setting at 360M, across loss type and chunk size. The native-script baseline is 0.9%.

layers). The auxiliary loss is weighted equally against the cross-entropy objective and applied to 60%. As with mixing, FKD documents are excluded from the alignment set.

We evaluate both an InfoNCE contrastive loss (van den Oord et al., 2018; Chi et al., 2021) and an L2 loss. L2 lacks negative sampling and is in principle susceptible to representational collapse; we hypothesise that the primary language-modeling objective provides sufficient regularization to prevent this.

Chunked matches. Since only document-level parallel data is available across our English–Arabic setup, we utilize an assumption of structural similarity between paired documents — chunks from similar relative positions in the document carry similar semantics. We choose a unit length n, split the language-A document into chunks of n words, and split the language-B document into the same number of equal-length chunks. The choice of n presents a trade-off: large n gives high confidence that paired chunks correspond but poor granularity, while small n gives fine granularity at the risk of misalignment. The structural-similarity assumption is somewhat defensible for machine-translated corpora such as ours due to the structural biases of machine translation.

Results. We sweep the chunk size n over $\{4, 20, 100, 250\}$ words for both losses (Table F.2). No configuration offers substantial benefit to knowledge generalization: CL_{eq} ranges from 0.3% to 2.5% across the sweep, against a 0.9% baseline.

G Soft Mapping

Several experiments use *soft mapping*: a single mechanism for partially sharing representations between two token sets, applied in two settings. In this appendix, we define it properly, describe each application, and present detailed results in Tab. G.1.

G.1 Mechanism

Let d be the embedding dimension of our model and $\mathcal{M}$ be a correspondence between two token sets (e.g., English_2 token and its English_1 counterpart). Under soft mapping with sharing rate p, each pair $(t, t') \in \mathcal{M}$ shares the first $\lfloor p \cdot d \rfloor$ embedding dimensions as tied parameters updated by gradients from both tokens; the remaining dimensions are independent. Thus $p = 0$ recovers disjoint embeddings and $p = 1$ recovers full mapping. Tokens without a counterpart in $\mathcal{M}$ are unaffected.

G.2 Diagnostic Use: English_1–English_2

Here soft mapping is purely a measurement instrument. The two languages are structurally identical and $\mathcal{M}$ is the exact shift correspondence between a token and its copy, so p controls the sole difference

		CL_{eq}	
p	**Free dims**	English_1–English_2	English–WWT-Ar
0%	960	-2.6%	—
20%	768	8.6%	1.0%
50%	480	82.7%	1.7%
90%	96	95.8%	9.6%
99%	10	99.1%	10.5%
99.9%	1	99.0%	10.2%
100%	0	100%	12.6%

Table G.1: CL_{eq} as a function of the sharing rate p in both settings. *Free dims* is the number of language-specific embedding dimensions remaining at each rate. for English_1–English_2 $p = 1$ corresponds to completely tied language copies — $\text{CL}_{\text{eq}} = 100\%$ by definition; for English–WWT-Ar it corresponds to standard (not soft) WWT.

between them. Sweeping p answers how much sharing is required before generalization emerges, rather than merely whether it does (§5.3). The sweep also calibrates the upper end of the CL_{eq} scale, since $p \to 1$ approaches perfect sharing by construction.

G.3 Practical Use: Soft WWT-Ar

Under full mapping, WWT-Ar tokens are indistinguishable from their English counterparts. This is what enables knowledge sharing, but it leaves the model no mechanism to represent which language it is generating: the output language must be fixed externally and code-switched generation is unsupported. Soft mapping resolves this—each language retains distinct token identities and shares only embedding dimensions. We do not evaluate code-switching ability in generation directly; we report only that the mechanism exists.

We additionally hypothesized that reserving language-specific dimensions would *improve* transfer, on the grounds that translation equivalents differ in connotation and possible additional meanings and that merging them discards information the model needs. We find no evidence for this: transfer increases with p (§6.3). Soft mapping is therefore best understood as trading a small amount of transfer for the practical properties above, rather than as an improvement over full mapping.

Construction. Because WWT can change the token count and does not provide a strict one-to-one token correspondence, Soft WWT-Ar requires an additional step: we first perform the WWT while shifting the resulting tokens by V into a disjoint range (as in English_1–English_2 we double the model vocabulary size to accommodate this), then tie a fraction p of each WWT-Ar token's dimensions to the corresponding English token.

H Dictionary Curation and WWT-Ar Construction

This appendix describes how we generate the large-scale word mapping dictionaries used for WWT operation in §6.1. We also provide more information about the costs of this mapping and where they originate.

H.1 Dictionary Construction

We parse the first 7M documents of our Arabic corpus and collect all unique words, then translate them with Qwen2.5-32B-Instruct (Qwen et al., 2025). The result is a dictionary with 11M entries. Any existing bilingual dictionary could be substituted, avoiding this step entirely. For Russian, we curate a similar dictionary following the same procedure on the Russian corpora. We execute the WWT operation (§6.1) on Arabic and Russian using these dictionaries

Index	Arabic	Russian
bare (none)	81.92%	55.78%
1–9	17.79%	38.82%
10–99	0.28%	5.30%
100+	0.01%	0.10%

Table H.1: Share of word replacements taking each conflict-index length.

Language	Tokens / doc.	Δ
Arabic	709	—
WWT-Ar	921	+29.9%
WWT-Ar excluding conflict IDs	834	+17.6%
Russian	1190	—
WWT-Ru	1459	+22.6%
WWT-Ru excluding conflict IDs	1165	−2.1%

Table H.2: Mean document length in tokens under the respective English–Arabic and English–Russian BPE tokenizers; deltas are reported relative to the native script tokenization. Word-wise mapping increases sequence length, more so for Russian than for Arabic. A substantial part of the fertility increase originates in the appended IDs used for conflict resolution.

H.2 Conflict Resolution

The mapping must be reversible. Where multiple Arabic words translate to the same English word, we differentiate them by appending an index: if two Arabic words both map to *dog*, one is remapped to *dog1*. This guarantees that every distinct Arabic word form receives a distinct WWT-Ar form, so no information is lost under the mapping.

Translation is many-to-one at scale — our dictionaries hold 11.05M Arabic and 13.23M Russian entries mapping onto 6.51M and 6.99M distinct English targets — but conflicted targets are not uniformly frequent in text, so most replacements need no index. In the Arabic corpus 81.9% of word replacements take the bare form and 99.7% need at most a single-digit index; Russian is heavier at 55.8% and 94.6% (Table H.1), which we attribute to its richer inflectional morphology, where case and number variants that are distinct surface forms frequently share a single English lemma.

H.3 Tokenization Fertility

Word-wise mapping does not preserve sequence length. Because dictionary targets and transliterated forms tokenize differently from their source words, mapped documents are longer: WWT-Ar documents average 921 tokens against 709 for native Arabic, an increase of 29.9%. WWT-Ru averages 1459 tokens against 1190, an increase of 22.6% (Table H.2).

Conflict resolution indices are the dominant source of WWT's sequence-length overhead, since they turn a word that would tokenize cleanly into a word-plus-digit token sequence. Excluding them, WWT-Ar's fertility penalty falls from $+29.9\%$ to $+17.6\%$ and WWT-Ru's from $+22.6\%$ to -2.1% (Table H.2). The overhead is therefore a property of our conflict-resolution policy rather than an inherent property of word-wise mapping itself.

H.4 Out-of-Dictionary Words

Words absent from the dictionary are handled by reversible transliteration - a character level 1-to-1 mapping into English script. for Arabic we use Buckwalter transliteration and for Russian we use GOST. We measure that in the Arabic corpus 99.7% of word occurrences are matched rather than transliterated (78% of unique words), and for Russian that number stands at 99.8% (92.7% of unique words).

This carries a genuine inference-time cost: generating a given amount of content under WWT requires proportionally more forward passes than in the native script.

It also has a consequence for how our results should be read. Because we hold the token budget fixed rather than the document count, the WWT-Ar and WWT-Ru models see fewer documents than their native-script counterparts—roughly 23% fewer for WWT-Ar — and therefore cover less of the foreign-language corpus. The mapped models nonetheless achieve comparable or better English perplexity (§6).

One might attribute that gain to reduced interference: with less foreign content competing for capacity, English would benefit. Our results do not support this explanation. English-only pretraining yields *worse* English perplexity than in a bilingual configuration (20.51 against 19.53 and 18.99; §4), so additional foreign content helps rather than competes at this scale, and the models are not capacity-limited. Reducing Arabic coverage should therefore reduce the benefit English receives, not increase it. That WWT-Ar delivers a larger benefit while contributing fewer documents indicates it contributes more per token—consistent with the increase in knowledge generalization characterized through the CL_{eq} score.

I Embedding-Space Analysis

Models fail to generalize knowledge between two identical copies of English (§5.2), yet both copies carry the same language and data distribution, so we should expect the model to learn the same structure for each. To test this, we compare the two halves of each model's embedding matrix and ask whether they are structurally similar, whether one linear map relates them, and whether matching tokens are placed together.

I.1 Setup

Corresponding tokens in the two English_1–English_2 vocabularies are directly aligned by row in the embedding matrix. To reduce noise from rarely updated rows, we retain tokens with at least $1{,}000$ expected exposures per language, estimated from corpus frequency and training budget. This leaves $36{,}170$ rows at 360M and $37{,}062$ at 7B, out of $65{,}536$ per vocabulary. To prevent a shared mean direction from inflating similarity, we normalize rows, mean-center each matrix, and renormalize.

We report four measures: *matched cosine*, the mean cosine between corresponding rows; *linear CKA* (Kornblith et al., 2019) for global structure; *mutual k-NN* (Huh et al., 2024) at $k = 10$ for local structure; and *linear-map cosine*, the matched cosine on a held-out 20% after fitting a least-squares map on the remaining 80%.

Within-model comparisons use all eligible rows at each scale; across-model comparisons use the shared eligible set. We analyze the tied embedding matrix at 360M and the LM head at 7B.[2] Following Gröger et al. (2026), we calibrate each measure against $K = 200$ random row permutations at $\alpha = 0.05$. On the resulting 0–1 scale, higher means more similar: 0 marks the permutation baseline and 1 identical spaces. Every within-model score exceeds all permutations ($p = 0.005$).

Intervention	CL_{eq}	Matched cosine	Linear CKA	Mutual k-NN	Mapped cosine
360M					
Baseline	-2.6%	0.012	0.682	0.654	0.560
Code-switching (60%)	71.8%	0.691	0.841	0.740	0.755
Same initialization	78.6%	0.939	0.957	0.854	0.960
7B					
Baseline	-1.5%	0.013	0.611	0.819	0.506

Table I.1: Null-calibrated embedding similarity and CL_{eq} across English_1–English_2 interventions. The baselines learn similar, linearly alignable spaces but show negligible direct token correspondence and no knowledge transfer; code-switching and same initialization bridge the spaces and transfer knowledge.

[2]The 360M architecture ties its input embedding and output projection; the 7B architecture separates them. We use the 7B LM head because it receives the vocabulary-wide prediction updates that also train the tied 360M matrix, whereas its input embedding remains close to initialization.

I.2 Results

Within models. At 360M, the baseline already has strong global and local structural similarity (Table I.1). Direct token correspondence is nevertheless negligible (matched cosine 0.012), while mapped cosine reaches 0.560 on held-out tokens. A single linear map therefore recovers substantial correspondence absent from the raw space. The 7B baseline repeats this pattern.

We further analyze the embeddings of two models that achieve high knowledge transfer. The model trained with the same initialization for corresponding tokens (§5.4) and substantial code-switching in 60% of the documents, which transfers well in the English$_1$–English$_2$ setting, unlike the more realistic English–Arabic (§4.2). In both models, we observe a sharp rise in matched cosine similarity alongside the jump in $\mathrm{CL}_{\mathrm{eq}}$ (Table I.1). Global and local similarities also increase from already strong baselines, indicating that the main change is direct token alignment.

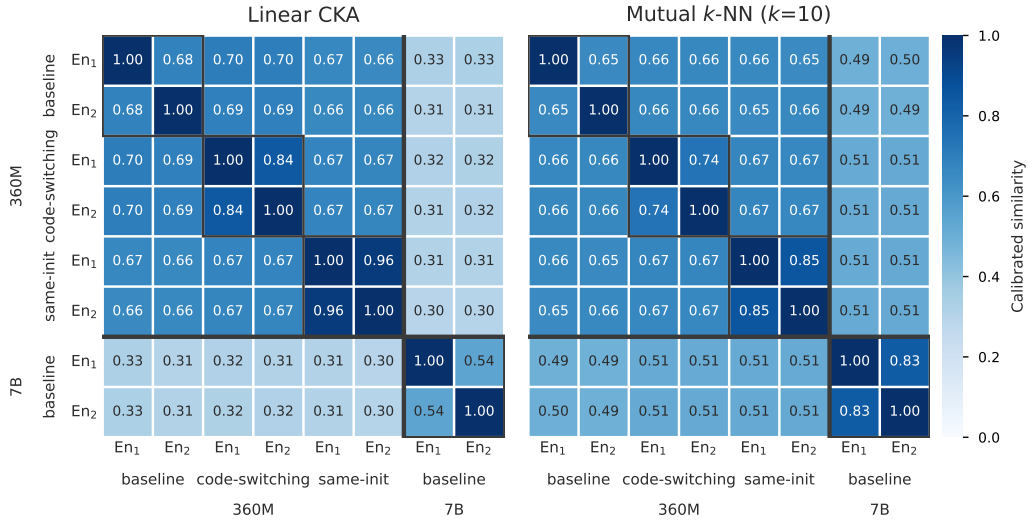

Figure I.1: Pairwise null-calibrated global (linear CKA; left) and local (mutual k-NN; right) embedding similarity. All models learn a common English geometry; code-switching and same initialization additionally bridge its two tokenized copies, yielding higher within-model similarity.

Across models. At 360M, the baseline's two halves fall within the same CKA and mutual k-NN ranges as cross-model pairs (Figure I.1). The baseline is thus no more structurally similar within one run than across runs: all 360M spaces learn a common English geometry. Models that generalize add a stronger within-model relation. Across scales, global and local similarities are lower than within-scale comparisons but remain substantial and well above the permutation baseline, despite differences in width and training budget. Combined with high matched cosine in the transfer conditions (Table I.1), this shows that the interventions bridge corresponding tokens within a common geometry; shared geometry alone does not yield knowledge transfer.

J Token Overlap and Anchored Arabic

Languages that share a writing system naturally exhibit partial token overlap through shared subwords, and this overlap correlates with better cross-lingual generalization (Ifergan et al., 2025; Kallini et al., 2025). This appendix examines whether recreating that overlap artificially induces knowledge sharing. It supports two claims in the main text: that the effectiveness of shared tokens depends on their frequency, not merely their number (§5.3), and that partial token-space overlap between natural languages does not deliver the gains of full merging (§6.2).

J.1 Controlled Overlap for English$_1$–English$_2$

We first sweep token overlap in English$_1$–English$_2$, where a subset of tokens is shared outright between the two copies while the remainder stay disjoint. Generalization rises steeply with the shared subset: sharing the 2% most frequent tokens already yields $\mathrm{CL}_{\mathrm{eq}} = 59.9\%$, reaching 89.0% at 64% overlap (Table J.1). Sharing a subset of tokens completely is thus far more effective than sharing all tokens partially — 2% overlap outperforms 20% soft-mapping by a wide margin (App. G).

Vocabulary share alone, however, does not determine the outcome. We repeat the 8% condition with a different token subset — one covering 47% of running text rather than 69.1% — and $\mathrm{CL}_{\mathrm{eq}}$ falls from 66.9% to 51.7%. The two conditions share the same fraction of the vocabulary and differ only in how frequently the shared tokens appear, so generalization tracks the *coverage* of the shared tokens rather than their count. (This second subset is the anchor set of the AnAr mapping introduced in §J.2.)

J.2 Anchored Arabic

Following these results we construct AnAr, a mapping that replaces only those Arabic tokens with a direct token-to-token match in our dictionary, leaving the rest in Arabic script (Figure J.1). AnAr targets the same barrier as WWT-Ar but far less aggressively, producing partial token overlap rather than a fully merged token space. Practically, AnAr preserves sequence length, avoiding the fertility overhead of WWT-Ar.

Language	Text
Arabic (original)	استخدم الولد الصغير قاموسا
AnAr	used الولد الصغير dictionary
WWT-Ar	used the boy the small dictionary
English (translation)	The small boy used a dictionary

Figure J.1: An Arabic sentence under each representation. AnAr replaces only words with a direct token-to-token dictionary match, leaving the rest in Arabic script; WWT-Ar maps every word. Both preserve Arabic word order (notice native Arabic is written from right to left) and grammar, unlike the English translation.

AnAr nonetheless fails to deliver meaningful knowledge generalization, with $\mathrm{CL}_{\mathrm{eq}} = 1.6\%$ against the WWT-Ar ceiling of 12.6%. The controlled sweep explains why: AnAr's anchors comprise 8% of the vocabulary but cover only 47% of English and 34.6% of Arabic running text, this low coverage, as we established, weakens the effect of partial vocabulary sharing (App. J.1), partially explaining the gap between effectiveness here and in the English$_1$–English$_2$ setup, more linguistic aspects may be in play.

K Language Modeling under WWT

The main text reports that WWT improves English perplexity (§6.2). This appendix provides the corresponding non-English measurements, so that the gain can be read against what it costs on the mapped side, and reports the monolingual controls that isolate how much each representation extracts from the English half of the corpus (Tab. K.1).

K.1 Measurement

We report bits per byte (based on Arabic script byte count) — BPB — rather than perplexity. WWT-Ar and native-script Arabic differ in token fertility (App. H.3), and a per-token measure would mechanically favour the higher-fertility representation: each WWT-Ar token carries less of the underlying text and is correspondingly easier to predict.

K.2 Arabic-Side Modeling

In the bilingual setting, mapping Arabic to WWT-Ar costs little on the Arabic side: at 360M parameters bits per byte rises from 0.7129 to 0.7160, a relative degredation of less the 0.5%. similarly, at 7B parameter scale BPB goesfrom 0.5679 to 0.5741, degradation of 1.1%. This is a small cost for the $14\times$ improvement in $\mathrm{CL}_{\mathrm{eq}}$ reported in §6.2, and it is achieved while the WWT-Ar model is exposed to roughly 23% fewer Arabic documents under our fixed token budget (§6.1).

Overlap	Coverage	CL_{eq}
$English_1 - English_2$		
0%	0%	−2.6%
0.1%	11.4%	−1.3%
0.5%	29.0%	43.1%
2%	50.1%	59.9%
4%	58.5%	61.2%
8%	**69.1%**	**66.9%**
8%	**47%**	**51.7%**
16%	79.2%	68.2%
32%	88.5%	77.2%
64%	96.1%	89.0%
English − Arabic		
0%	0%	0.9%
English − AnAr		
8%	47%/34.6%	1.6%
English − WWT-Ar		
100%	100%	12.6%

Table J.1: Knowledge generalization as a function of token overlap. *Overlap* is the share of vocabulary types shared between the two languages, and *coverage* is the share of training tokens that they account for, reported per side where the two differ. The controlled setting achieves far higher CL_{eq} than the natural one, while within each setting CL_{eq} increases with overlap. These experiments were performed at the 360M parameter scale.

Representation	Monolingual	Bilingual	Δ
Native Arabic	0.7155	0.7129	0.0026
WWT-Ar	0.7344	0.7160	0.0184
Native Russian	0.5892	0.5763	0.0129
WWT-Ru	0.6164	0.5870	0.0294

Table K.1: Non-English language modeling in bits per byte (BPB), normalized by native-script bytes throughout (§K.1); lower is better. *Bilingual* models are trained on a 50/50 token split with English, *monolingual* models on the non-English half alone. Δ is the improvement contributed by the English half. Experiments performed at the 360M parameter scale.

K.3 Monolingual Controls

Bilingual numbers alone cannot say how much of each model's Arabic-side performance comes from the English half of the corpus. We therefore train monolingual counterparts for both representations and compare each bilingual model against its monolingual baseline (Table K.1), trained only on the Arabic half of the data (30% less data for WWT-Ar). The quantity of interest is the *improvement* from adding English: if the unified token space allows the model to draw more from English, WWT-Ar should show the larger improvement. Indeed, we find that the monolingual Arabic model outperforms the WWT-Ar model with BPB of 0.7155 compared to 0.7344 (2.6% difference), however, the WWT-Ar model gains much more from the cross-lingual data — a Δ of 0.0184 compared to 0.0026, roughly $7\times$ matching directional CL_{eq} measurement gain (Tab. K.1).

L Shuffled-Mapping Controls

Our WWT operation is overwhelmingly a translation rather than a transliteration operation (§6.1), meaning it maps Arabic tokens or token sequences to their semantically matching English counterparts. This appendix provides implementation details for the experiments performed in §6.4 to determine the importance of this property as opposed to plain token superposition. We run three controls that preserve token sharing while destroying correspondence, and find they all fail. We perform these experiments at the 360M parameter scale.

L.1 Shuffled WWT Dictionary

We destroy semantic correspondences by shuffling the keys and targets of our large-scale dictionaries used for WWT, then proceeding with the WWT process as usual, mapping words to their shuffled dictionary mapping and transliterating out-of-vocabulary entries. We find this process hurts knowledge generalization between the two languages compared to our unshuffled WWT baseline, with English–WWT-Ar CL_{eq} falling from 12.6% to 3%, and English–WWT-Ru falling from 23.3% to 3.6%.

L.2 Transliteration

Transliteration-only mapping provides a similar effect: a unified token space, but tokens aren't mapped to their semantic counterparts. We perform the same reversible transliteration used for WWT but apply it to all words. Essentially applying WWT with an empty dictionary. We find that transliterating Arabic does not help with resolving compartmentalization between it and English, with a CL_{eq} score of -0.3%.

L.3 Permuted English_1–English_2

We additionally pretrain a model on two language clones — English_1–English_2 — in which both languages occupy a single shared token space, but the token identities of one copy are permuted: $\sigma : \{0, \ldots, V-1\} \to \{0, \ldots, V-1\}$. Formally, while a text sequence is tokenized in English_1 as an array of token IDs $(t_1, t_2, \ldots, t_n)$, English_2 tokenizes the parallel sequence into $(\sigma(t_1), \sigma(t_2), \ldots, \sigma(t_n))$, both share the same $\{0, \ldots, V-1\}$ token space

This provides a fully controlled environment where the entire token inventory is shared and the two languages remain structurally identical, with the only barrier to generalization being the mismatch between token idetities. Yet, we observe no knowledge generalization with a CL_{eq} score of -2.6% — same as completely disjoint token spaces.

L.4 Interpretation

Shared tokens transfer knowledge only when they carry corresponding meanings. Resolving disjoint tokenization is therefore not a matter of token superposition, and a semantically faithful mapping is necessary. The transliteration control is notable, as previous work has suggested transliteration as a path to unlocking multilingual ability (Moosa et al., 2023; J et al., 2024; Saji et al., 2025; Jayakumar et al., 2026), we note that successful examples in past literature focus mostly on related languages where cognates are common and therefore transliteration can supply the semantic matches we find necessary.

Setting	Token space	Semantic mapping	CL_{eq}
English − Arabic			
Native Arabic (baseline)	disjoint	—	0.9%
WWT-Ar	shared	yes	12.6%
Shuffled WWT-Ar	shared	no	3.0%
Transliterated Arabic	shared	no	−0.3%
English − Russian			
Native Russian (baseline)	disjoint	—	1.9%
WWT-Ru	shared	yes	23.3%
Shuffled WWT-Ru	shared	no	3.6%
$English_1$ − $English_2$			
Disjoint copies (baseline)	disjoint	—	−2.6%
Fully shared copies	shared	yes	100%
Permuted copies	shared	no	−2.6%

Table L.1: Semantic controls at 360M. Each block pairs a semantically faithful shared token space against controls that share tokens without sharing meanings. Sharing a token inventory without semantic correspondence performs at or near the disjoint-token baseline in every pair, while the semantically faithful mapping is the only condition that improves generalization. for $English_1$–$English_2$ Fully shared copies are at CL_{eq} of 100% by definition