

# Beyond the Rosetta Stone: Unification Forces in Generalization Dynamics

Carter Blum\*, Katja Filippova\* & Ann Yuan\*

Asma Ghandeharioun, Julian Zimmert, Fred Zhang & Jessica Hoffmann

Mor Geva<sup>†</sup>, Martin Wattenberg<sup>‡</sup>, Tal Linzen<sup>§</sup> & Lucas Dixon

Google DeepMind

{carterblum, katjaf, annyuan}@google.com

## Abstract

Large language models (LLMs) struggle with cross-lingual knowledge transfer: they sometimes hallucinate when asked in one language about facts expressed in a different language during training. This work introduces a controlled setting to study the causes and training dynamics of this phenomenon by training small Transformer models from scratch on synthetic multilingual datasets. Depending on (1) the correlation between facts and the language they were learned in (informativeness), and (2) the ease of language identification (extractability), models either develop unified representations across languages or separate representations; only when representations are unified do facts transfer across languages. Based on these insights, we propose a unifying perspective which explains a range of prior observations concerning cross-lingual transfer in multilingual LLMs. Our work shows controlled settings can shed light on pre-training dynamics and suggests methods to encourage representational unification as part of training that would improve LLMs’ cross-lingual transfer.

## 1 Introduction

Language models are known to hallucinate facts, especially during cross-lingual factual recall—that is, when facts are learned in one language, but recalled in another language (Goldman et al., 2025). In this work, we are interested in understanding what gives rise to these cross-lingual hallucinations and how to mitigate them. Therefore, we focus our analysis on the learning dynamics during pretraining (Biderman et al., 2026), and particularly on the impact of training data properties and the resulting model’s representations.

We investigate the connections between data properties, models’ representations after learning from this data, and final downstream cross-lingual performance. To do so, we need to 1) have tight control on the training data properties (we want to be able to vary each parameter individually to identify the relevant ones, and causally intervene on some of them to test hypotheses) and 2) analyze many models trained on these different datasets, and the subsequent internal representations. For these reasons, we use a “Petri dish” methodology: we train small transformers from scratch on synthetic datasets while systematically and methodically varying their distributional properties. Such synthetic setups and training data interventions are well-established in the literature, and has already enabled crucial insights into model capabilities (Chan et al., 2022; Park et al., 2024; Allen-Zhu, 2024).

Thanks to this setup, we obtain a plethora of models which have perfect<sup>1</sup> factual recall about facts expressed and recovered in only one language, but a wide range of cross-lingual factual recall accuracy (from 40% to 100%). We consequently identify a pivotal early phase where a model develops either unified or separate representations for identical facts across languages.

---

\*Equal contribution. <sup>†</sup>Also at Tel Aviv University. <sup>‡</sup>Also at Harvard University. <sup>§</sup>Also at NYU.

<sup>1</sup>Here, ‘perfect factual recall’ indicates that the in-language test accuracy consistently reaches >99% after convergence, as demonstrated in Figure 2a and the training curves in Appendix Figure 11b.

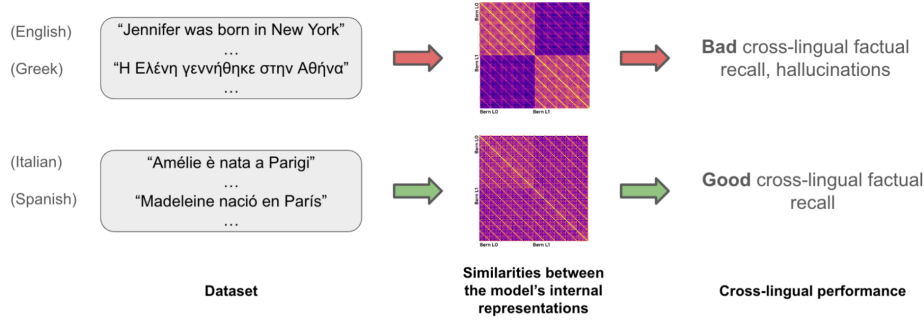


Figure 1: Learning dynamics for cross-lingual factual recall. On top, the languages are informative<sup>2</sup>—aka predictive of content (here, the city)—and extractable—Greek and English are very different. This leads the model to learn separate representation for each language, which then results in poor cross-lingual downstream performance. Inversely, at the bottom, the languages are not as informative (city not correlated with language) nor extractable (Italian is close to Spanish). This leads the model to learn unified representations, which then results in good cross-lingual factual recall.

A fully-pretrained model has good cross-lingual factual recall if and only if it develops early unified representations. To capture this effect, we introduce a new metric, the language unification score, and we show that its value in these early steps of training is predictive of cross-lingual generalization in the fully trained model. This score is computed over training examples, and is as predictive of cross-lingual model performance as a validation set would be, but without the need for a held-out set.

We then examine the properties of the data that lead to separate representations, and therefore to subpar cross-lingual generalization. Intuitively, the representations will separate if we learn how to distinguish the languages before we learn the facts. This is more likely to happen in two cases: if the language is **informative**, that is, if knowing the language helps with learning the facts (e.g. there are more Spanish cities in texts in Spanish), and if the language is **extractable**, meaning if it is easy to recognize from which language each token/sentence comes (e.g. in Greek, which has a unique script).

If we increase the amount of parallel data (facts expressed in both languages), we reduce the informativeness of both languages, which in turn increases generalization—a well-known fact. Similarly, our findings provide a unifying perspective on multiple other seemingly disparate observations from prior work about cross-lingual transfer in LLMs. In summary, our core contributions are:

1. We systematically vary parameters to generate synthetic data in **synthetic languages** with different distributional properties. We train models on each of them and compare in-language and cross-lingual factual recall. By inspecting their internal representations, we identify a crucial phenomenon: the models perform well if and only if they develop **unified representations** across languages early in training.
2. We show we can measure this unification via the **language unification score**. This score is computable during early training, and is strongly predictive of downstream cross-lingual generalization. It is as effective for picking the best models as using a validation set.
3. We identify the leading causes of representation unification failure: whether the languages are **informative** (predictive of content) and **extractable** (easily differentiable from one another). We then test this hypothesis by artificially increasing and decreasing how informative and extractable the synthetic languages are and thus confirm our findings. This dual lens provides a unified explanation of how script, vocabulary size and embeddings have been shown in seemingly disparate previous works to affect cross-lingual knowledge transfer.

<sup>2</sup>From top to bottom, aside from the first example, the sentences translate to (Greek) "Elini was born in Athens", (Italian) "Amélie was born in Paris", (Spanish) "Madeleine was born in Paris".

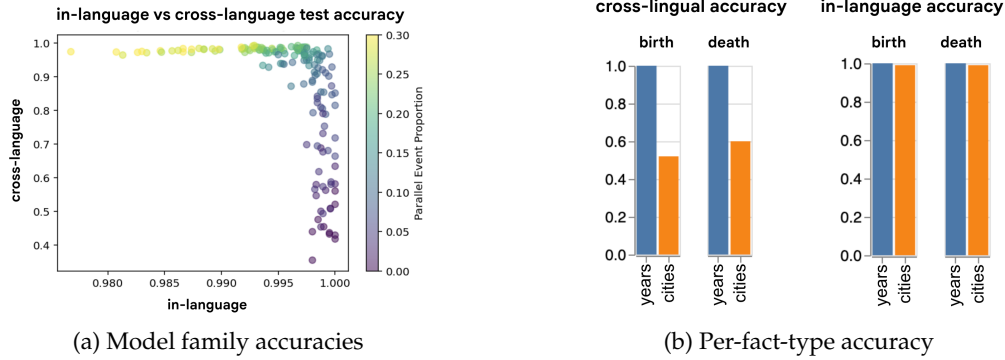


Figure 2: **Left:** In-language vs. cross-language test label probability across Pythia models (pretrained on datasets expressing the same facts) across parallel data ratios. In-language performance does not predict cross-lingual performance. **Right:** Accuracies for in-language and cross-language evaluation of one model for the four fact types. Model attains perfect cross-lingual accuracy when predicting years but not cities, despite perfect in-language accuracy for that task.

## 2 Methodology

**Data creation and training/test sets** In our tiny setting, we consider only 4 types of facts: *A person is born in a city* (birth-city), *a person is born during a year* (birth-year), *a person died in a city* (death-city), and *a person died during a year* (death-year). During our experiments, we use **synthetic languages**. However, for clarity, the examples in the paper are in English or Spanish. For instance, Alice Brown being born in Berlin is a fact of type birth-city. We create thousands of such facts, and we express them in templates. The potential templates are the same for each fact, and they are arbitrarily split into two sets representing each "language". By default, the templates do not have overlapping tokens across the two distinct languages; however, multiple templates within the same language may share tokens. The fact arguments (e.g., *persons*, *cities*, *years*) remain shared across both languages. Our datasets are composed of these facts expressed in these templates, e.g. "Alice Brown was born in Berlin" or "Alice Brown's life started in Berlin" or "Alice Brown nació en Berlin" are all valid pieces of data. For each fact, we decide if it is parallel or unparallel data. If it is parallel data, we express this fact in all the templates of both languages, and this data becomes part of the training set. For unparallel data, we pick only one of the two languages, and then select some templates (leaving at least 1 out, similarly to Allen-Zhu (2024)). The fact expressed in all the templates selected (and therefore in only one language) joins the training set, while the left-out templates in the same language join the **in-language** test set. The same fact expressed in all the templates of the other language become part of the **cross-language** test set. In addition to the overall in- (or cross-) language accuracy, for some experiments we report accuracy per fact type. For more details about the generation procedure and the exact templates, see Appendix 9.1.

Intuitively, increasing the amount of parallel data should improve generalization across languages. To study different regimes, we vary the proportion of cross-lingual events and generate multiple datasets for different ratios, all containing the same facts.

**Tiny model training and evaluation** We train small Transformer models using standard configurations (Pythia, Gemma 2) from random initialization on our synthetic datasets for multiple epochs, monitoring the in- and cross-language accuracies during training. The total parameter count in our models is typically around 2M (six layers, four attention heads, and a hidden size of 128). As described previously, we vary the following parameters when creating a training set: (1) amount of parallel data, and (2) the discrepancy in entity frequency between languages. We report the fraction of parallel data by event, from 0% (no parallel data) up to 30% with a 2% step size. See App. 9.2 for more details.

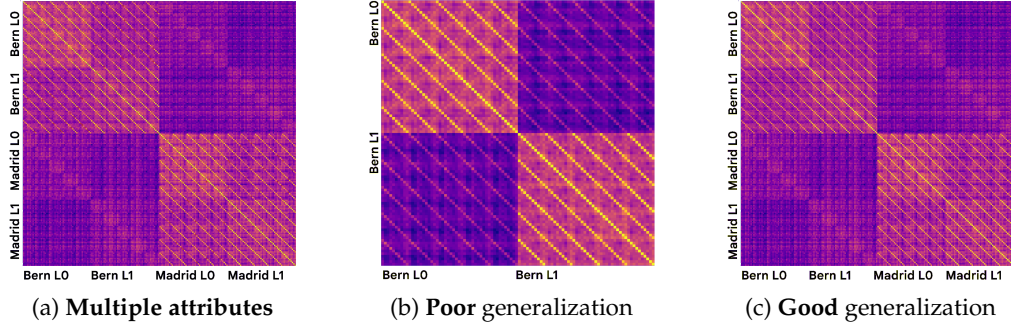


Figure 3: Generalization patterns across different experimental setups.

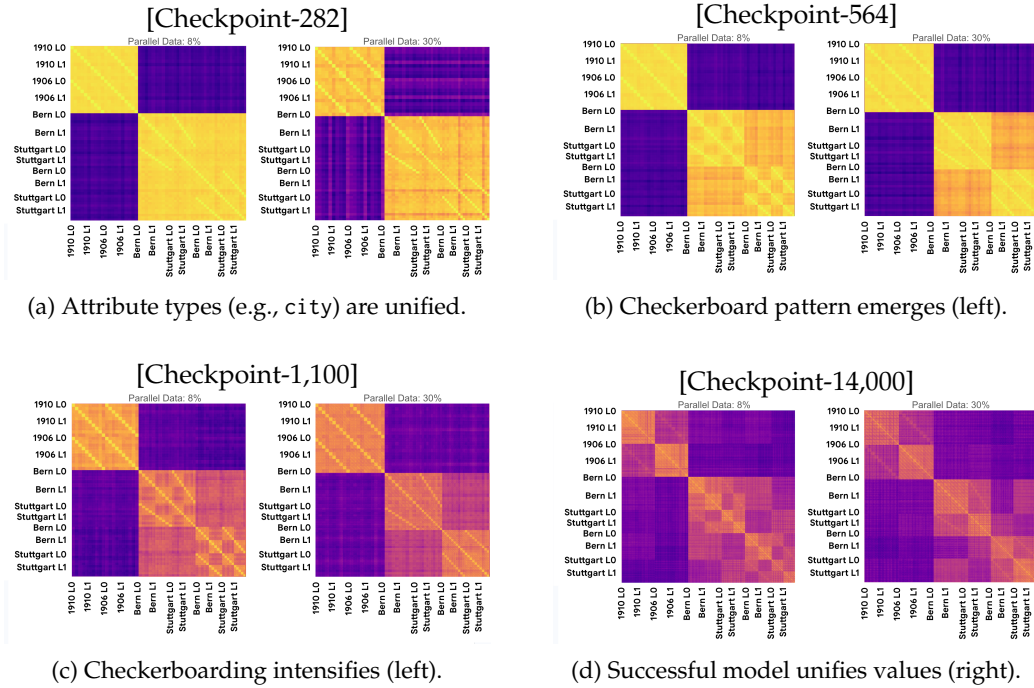


Figure 4: Pairwise similarity matrices between activation-based representations at the token preceding the attribute across checkpoints. One model (on the right in every pair of matrices) eventually achieves 100% cross-lingual test accuracy while the other (on the left) does not. Red means high similarity. Detailed descriptions for each stage are provided in the main text. Note: Labels such as ‘Bern’ and ‘Stuttgart’ are repeated on axes to demonstrate the model’s clustering of distinct fact types (e.g., birth versus death) for the same city.

### 3 Learning Stages

We analyze representations based on: (1) activations from the residual stream, and (2) gradients of model parameters. By monitoring cosine similarity between representations during training, we can pinpoint when semantically related inputs across languages converge and diverge. To obtain activation-based representations we take the residual stream contents for the token immediately preceding the attribute to be predicted (e.g. *Alice Brown was born in*). Unless otherwise noted, we concatenate the activations of each layer to form this representation. The use of such embeddings for the purpose of model analysis and visualization is well established (nostalgebraist, 2020; Ghandeharioun et al., 2024). To obtain gradient-based representations we use the model weights’ gradients at the attribute token. Intuitively, if the gradients between two examples are similar, then they exert a similar influence during training, and are processed by similar model parameters. We draw inspiration from methods for identifying influential training inputs by comparing the gradients of

training data to test data (Koh & Liang, 2017; Schioppa et al., 2023; Ruis et al., 2025), and leverage a recent, computationally efficient approximation developed by Chang et al. (2025).

Fig. 3a shows pairwise cosine similarities between activation-based representations<sup>3</sup> of a Gemma model trained on a 50-50 language split and 16% cross-lingual events. In these and following experiments, we adopt a 50-50 language split and a moderate percentage of cross-lingual events to simulate a balanced bilingual environment with a realistic fraction of parallel data. This setup allows us to isolate the effects of language unification without introducing confounding variables like dataset imbalance. Representations are computed from training examples and, while referring to distinct *birth-year* facts, all have *Bern* or *Madrid* as the (correctly predicted) birth-city attribute. Within each attribute, examples are grouped by language ( $L0, L1$ ). In this model with median generalization ability, examples with the same attribute (e.g., *Madrid*) are more similar to each other than to examples with a different attribute (*Bern*)—this is clearly visible in the two large red blocks. However, within the red matrix corresponding to the same attribute (e.g., *Madrid-Madrid*, bottom right), two sub-blocks are visible, indicating that examples of the same language are more similar than examples across languages. For comparison, Fig. 3b-3c show the similarities for a *single* birth-city attribute (*Bern*), again grouped by language, from two models—with worse and better cross-lingual factual recall (8% vs. 30% cross-lingual events). Only Fig. 3b has four distinct blocks, indicating cross-lingual dissimilarity.

When does separation by language happen in training? Consider Fig. 4, which shows pairwise similarity matrices for two models across checkpoints. One model (on the right in every pair of matrices) eventually achieves perfect cross-lingual transfer while the other (on the left) does not. The models’ training sets are identical with respect to factual content, languages used, language split (50-50) and number of examples, but differ in the amount of parallel data—30% events vs. 8%. Note that the trends on display are observed consistently across dozens of models trained with different languages, language proportions, model scales, etc. Each matrix in Fig. 4 shows pairwise similarities between 300 examples equally split between birth-year, birth-city, death-city, each corresponding to two values (1906, 1910 for birth-year, *Bern*, *Stuttgart* for birth-city, *Bern*, *Stuttgart* for death-city). Again, within each attribute value, examples are grouped by language ( $L0$  and  $L1$ ). We observe that poorly-generalizing models undergo a signature phase (around checkpoint-564) wherein language identity, rather than semantic equivalence, drives representational similarity.

We quantify our observations with a *language unification score*, which predicts cross-lingual generalization. Intuitively, the language unification score measures how much the model *avoids* the checkerboarding on the right side of Fig. 4.

## 4 Unification Predicts Cross-Lingual Performance

Concretely, the language unification score captures the similarity between semantically equivalent cross-lingual datapoints against a baseline of similarity between semantically distinct same-language datapoints. We define it as  $\text{Unification}(\theta, \mathcal{D}) := E_{X,Y \sim \text{Facts}(\mathcal{D})} [\text{sim}_\theta(X, Y) / \text{sim}_\theta(X, X)]$  where  $X, Y \in \text{Facts}(\mathcal{D})$  samples the datapoints corresponding to each fact in the dataset  $\mathcal{D}$ , grouped by language.  $X$  and  $Y$  are representations of the same fact in different languages (see Appendix Fig. 12 for pseudocode). For our experiments, we let  $\mathcal{D}$  be the parallel examples from the training set.  $\text{sim}_\theta$  is the average cosine similarity between the representations of both sets. The language unification score correlates strongly (Pearson’s correlation coefficient  $>0.95$ ) with cross-lingual accuracy, for both activation-based and gradient-based representations (Fig 5).

Not only does the language unification score correlate strongly with model quality, it can be used to *select* training runs that will generalize well. In fact, as Fig. 6 (left) demonstrates, we find that utilizing the language unification score can be as effective as collecting a small test set. These experiments were performed over 110 runs varying the fraction of parallel examples (from 0% to 20%), the ratio of monolingual examples in the majority vs

<sup>3</sup>Gradient-based plots are in Fig. 13 in the Appendix.

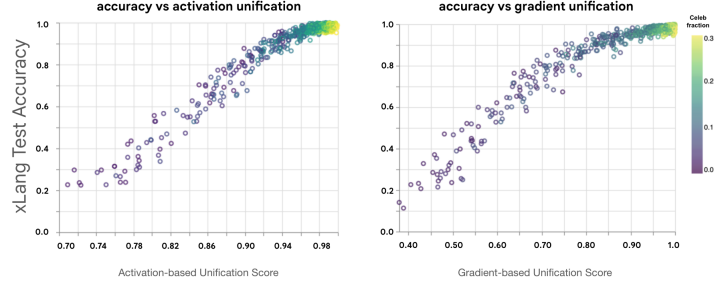


Figure 5: Activation language unification scores (left) and gradient language unification scores (right) correlate strongly with cross-lingual factual recall accuracy (0.97 and 0.94 PCC, respectively) (see Fig. 14 for layerwise results). Each datapoint corresponds to a different model training run. Note that the fraction of parallel examples (denoted by color) also correlates strongly with generalization ability.

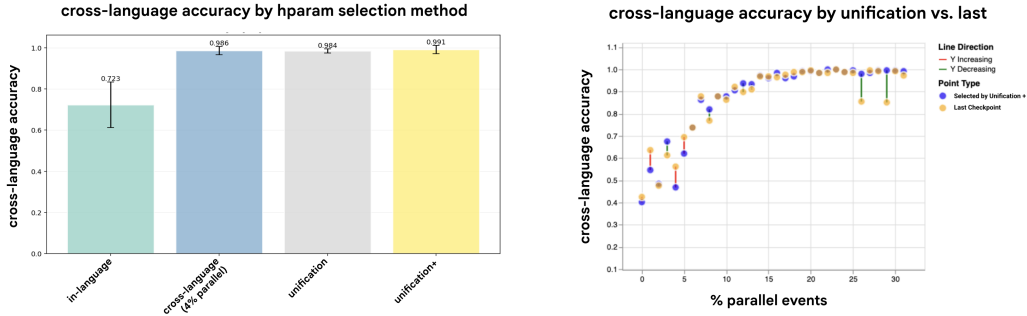


Figure 6: Comparison of hyperparameters and checkpoint selection strategies.

minority language (from 1:1 to 20:1) and the amount of token overlap within a language. We repeatedly sampled 33% of runs and compared the cross-lingual test performance of the runs chosen by a variety of heuristics. These heuristics include: in-Lang Test (selecting the model with the best in-language test performance), xLang Test (k%) (selecting the model based on a smaller cross-lingual test set, with k% of the cross-lingual test set being used for model selection. This is a strong baseline, as it effectively uses more data), and language unification score (last token language unification score).

We also evaluate language unification scores as a method for selecting a checkpoint, as it may be expensive or difficult to collect cross-lingual validation data. In Fig. 6 (right), we compare to a baseline that selects the last checkpoint among those with the highest in-language accuracy. While both methods tend to choose checkpoints with similar quality, selecting using language unification scores avoids some scenarios where the model begins overfitting on in-language recall. This represents an exciting new direction for model selection – mechanistic analysis of models can improve our ability to predict generalization.

## 5 Analyzing the Language Feature

As the previous analysis demonstrates, and in line with observations from analyses of LLMs, cross-lingual factual recall fails if the model’s internal representations are separated by language, that is, when *language identity is strongly encoded* in the hidden representations. We hypothesize that the strength of this encoding, or the language feature footprint, is determined by the **informativeness** of language identity for the prediction task and its **extractability** from the data. This is theoretically grounded: seminal work by Saxe et al. (2019) shows that models learn features in descending order of the variance they explain in the training data; e.g., a high-level *plant-vs-animal* distinction is learned before fine-grained categories like *bird-vs-fish*. Similarly, Lampinen et al. (2024) argue that easier-to-learn features tend to dominate a model’s internal representations and explain more variance than difficult-to-extract features. To restate our hypothesis:

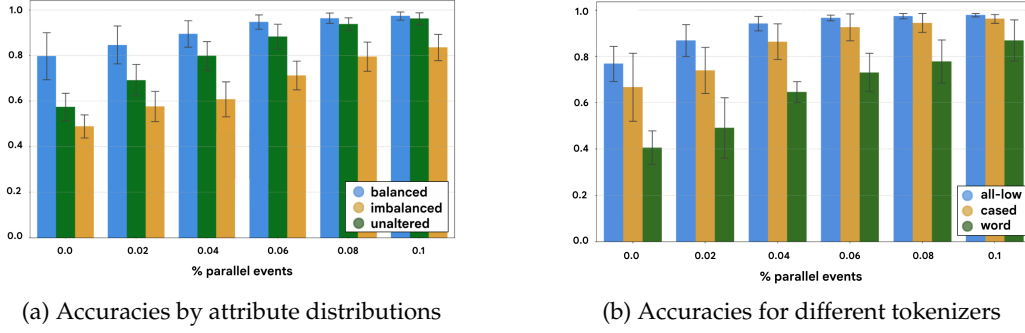


Figure 8: Cross-lingual accuracy for different values of parallel events for models (a) trained on unaltered, balanced and imbalanced datasets, (b) trained with the character (all-low and cased) and word tokenizers.

If the language identity provides a useful signal about the label (is **informative**) and is easy to recognize (is **extractable**), it will be learned early and dominate the representation. Conversely, if language identity is not informative or hard to extract, its representational footprint will be small and representational separation is less likely.

## 5.1 Language informativeness

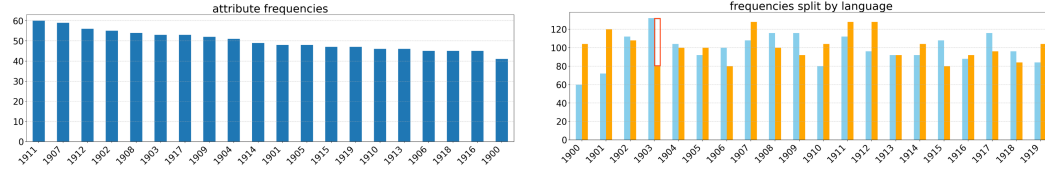


Figure 7: Frequencies of the birth-year attribute values in the Knowledge Graph (KG) (left) and in a dataset, split by language (right). Light blue and orange on the right represent the frequencies in the first and second languages respectively after template verbalization. The red rectangle highlights the difference in example counts between two languages for a particular year: it represents the number of examples for that attribute needing to be added to the second language (orange) to create a balanced dataset, versus being added to the first language (blue) to create an imbalanced dataset.

How can the language of an example be a useful signal for prediction? Strictly speaking, attributes (e.g., 1911) are a function of the subject entity (e.g., *John Smith*) and the fact type (*birth-year*). However, the distribution over attribute values is not uniform in our Knowledge Graph (KG) or datasets (Fig. 7). Thus, a spurious correlation (positive correlation) exists between the language identity and the predicted attribute. This property of our toy data mimics real multilingual datasets: for example, Spanish texts reference Spanish cities more often than German texts do. The language feature does therefore provide a useful prior over attribute values, helping to reduce loss early in training, *before* the model learns to recall facts by combining the subject and relation representations (Geva et al., 2023). To measure the role of the mutual information between the language and the attribute variables, we create two dataset versions from an existing base dataset with little ( $<10\%$ ) parallel data and with an equal split between languages: balanced includes additional examples (all corresponding to **new** events) to equalize the example count for every attribute value across the two languages, thereby *minimizing* the language feature’s informativeness without increasing the amount of parallel data; and imbalanced includes the same number of additional examples, created from the same set of new events, but adds them in the language in which the attribute is already present more frequently, amplifying the existing discrepancies and therefore *increasing* the language feature’s informativeness.

Fig. 8 (left) shows cross-lingual accuracies for the three settings. Cross-lingual generalization is consistently worst when attribute distributions are imbalanced and correlation between language and attribute is high (checkerboarding is correspondingly more prevalent in this

setting, see App. Fig. 16). This experiment reveals a clear correlation between the informativeness of the language feature and cross-lingual performance: cross-lingual accuracy improves even when the amount of facts expressed in both languages stays zero.

## 5.2 Language extractability

We next test the hypothesis that making the language feature easier to extract also boosts its influence on the model’s representations. In our default setup, the language feature is trivial to extract: our synthetic languages do not share vocabulary, and our tokenizer is word-based. Thus every token (except for entity arguments like names, cities, and years) is a perfect indicator of language identity. To make the language feature harder to extract, we switch to a *character-based* tokenizer while keeping all other aspects of model training unchanged. A character model needs to accumulate information over many tokens to recognize the language. Furthermore, to isolate the effect of language feature extractability from other potential effects of this tokenizer change, we contrast three settings: word Baseline word-based tokenizer and the original dataset (highly extractable language feature); all-low Char-based tokenizer, with all the templates in lowercase (less extractable); cased: Char-based tokenizer, where templates in the first language are in lowercase and the second are uppercase (more extractable language feature than char but less than word because people’s names are spelled normally, introducing lowercase letters).

The last two models are directly comparable, as both use the same character-based tokenizer on the same examples. The only difference is that the language feature is more easily extracted by the latter model, since the token sets used in the two languages are disjoint (upper-case vs lower-case characters). The results in Fig. 8 (right) confirm that when the language feature is more difficult to extract, it has a direct, positive effect on cross-lingual generalization: for the same percentage of parallel data, the char-all-low models, where the language feature is least extractable, consistently achieve the highest accuracy. Similarly, checkerboarding is more present in ‘cased’ models (App. Fig. 17)

We observe a similar effect when making language harder to extract by increasing the number of templates. We conduct experiments where there is no token-level indication of language, making it difficult for the model to correctly group all templates into languages. Increasing the number of templates substantially improves generalization, as demonstrated in App. 10.6, while neither increasing the number of events by 10x nor increasing the training duration by 10x improves generalization.

Recall that Figure 2a shows both that parallel data improves cross-lingual generalization, and that generalization can take place without parallel data. The experiments in this section suggest a new perspective on parallel data’s role in cross-lingual transfer: Increasing parallel data naturally reduces the language feature’s informativeness about attribute distributions.

## 6 Interpreting Existing Observations

Our key contribution is to shed light on why language models might fail to transfer factual knowledge cross-lingually. Here we show how our findings help explain various seemingly unrelated observations in prior work:

**Script similarity** Several studies have observed that cross-lingual performance correlates more strongly with script similarity than with geographical, linguistic, or cultural proximity. For instance, Greek—an Indo-European language—shows low correlation with other European languages (Liu et al., 2025). Conversely, Indonesian—an Austronesian language utilizing the Latin script—demonstrates better cross-lingual transfer with English than other Asian languages (Goldman et al., 2025). Further supporting this, research has shown that handling unseen languages (Muller et al., 2021a) or strengthening low-resource translation (Shurtz et al., 2025) heavily depends on mitigating script divergence.

While seemingly counter-intuitive, our framework offers a clear explanation: language identity is trivial to extract from a unique script (e.g. Greek). Consequently, the “language

feature" is absorbed early in training, leading to siloed representations. In contrast, shared scripts (like Latin) delay this extraction, allowing for better transfer.

**Aligning embeddings** Li et al. (2024) initialize embeddings for aligned words to be similar prior to pretraining to promote shared representations. Rather than directly replicating these interventions, our framework provides a mechanistic rationale for why they work: pre-alignment and increased vocabulary overlap obscure the language feature, reducing its extractability. This encourages the early development of the unified representations that we show are critical to cross-lingual transfer.

**Shared vocabulary** Patil et al. (2022) propose increasing vocabulary overlap between languages prior to training, demonstrating that this improves cross-lingual knowledge transfer. These empirical benefits of shared vocabulary were also highlighted in early studies of Multilingual BERT (Pires et al., 2019; K et al., 2020) and large-scale cross-lingual representation learning (Conneau et al., 2020a), which identified subword overlap as a key driver of zero-shot transfer. Following the same logic, (Shurtz et al., 2025) proposes mapping distinct scripts to shared phonemes, which dramatically improves cross-lingual generalization. We interpret this result similarly: increased vocabulary overlap makes the language identity feature harder to extract, thereby reducing its representational footprint and improving cross-lingual accuracy. Furthermore, Qi et al. (2023) find that vocabulary overlap correlates with cross-lingual factual consistency. They thus speculate that larger models do not necessarily show better cross-lingual consistency because their vocabularies are larger. We add another layer of interpretation: larger vocabularies simplify language identification, which in turn increases the prominence of spurious language features.

**Other work** Our findings also contextualize other observations in real LLMs. Mousi et al. (2024) note that deeper layers exhibit more cross-lingual alignment, mirroring our observation that language checkerboarding is most prominent in lower layers. Furthermore, while Wen-Yi & Mimno (2023) find that models like XLM-R cluster embeddings by script and mT5 by semantic similarity, our framework suggests these representational differences may arise from how distinct training objectives influence the informativeness and extractability of the language feature.

**Vocabulary Overlap in More Recent LLMs** To replicate our observations regarding the effects of tokenization (Sec. 5.2), we study vocabulary overlap between languages as a predictor of cross-lingual consistency, following Qi et al. (2023). We re-run the vocabulary similarity experiment by Qi et al. (2023) with **four recent open LLMs**: Gemma-3-4B (Team et al., 2025), Llama-3.2-3B (et al., 2024), Qwen3-4B (Yang et al., 2025), and Mistral-7B (Jiang et al., 2023). We use Flores-200 (Costa-Jussà et al., 2022) dataset that consists of 2,000 sentences translated from English into different languages. We run those sentences through the model tokenizer to get a model-specific vocabulary for each language. For each pair of languages  $A$  and  $B$  with vocabularies  $V_A$  and  $V_B$ , we calculate their Jaccard similarity ( $S = |V_A \cap V_B| / |V_A \cup V_B|$ ). We then group examples by their source and target language, and for a more intuitive analysis, focus on examples with the source language of English. We calculate the Pearson correlation between  $S$  and cross-lingual consistency across models (defined here as the consistency metric for factual queries introduced by Qi et al. (2023)) and observe an average coefficient of 0.69, ranging from 0.43 to 0.85 across different models (per-model details are in App. 10.7). This finding aligns with Qi et al. (2023) which reported a positive correlation between cross-lingual consistency and vocabulary similarity showing that vocabulary overlap explains a significant amount of variance in cross-lingual transfer.

## 7 Discussion and Related Work

Many studies look into causes of model hallucinations and attribute them to training and sampling noise, gaps in pretraining data (Xu et al., 2024), and misaligned incentives in post-training (Schulman, 2023). However, these fail to explain *cross-lingual* factual errors, nor does increasing model scale solve the problem (Aggarwal et al., 2025; Qi et al., 2023). Prior work

finds that LLMs may develop both a lingua franca for factual knowledge (typically based on English) and distinct language silos (Aggarwal et al., 2025; Lim et al., 2025b; Schut et al., 2025; Lim et al., 2025a, *inter alia*), and that their hidden representations can be language-agnostic or language-specific depending on the layer (Wang et al., 2025). This builds upon foundational work demonstrating that shared cross-lingual structures naturally emerge in pretrained models (Conneau et al., 2020b), and that successful zero-shot transfer requires representations to align across languages before predictive capabilities can transfer (Muller et al., 2021b). However, the root cause of these phenomena is not understood, as most studies treat models as static artifacts. Such analysis, while valuable, cannot explain how knowledge *arises* during training, and cannot lead to effective pretraining interventions. While some have investigated the training dynamics of knowledge acquisition in multilingual LLMs (Zeng et al., 2025; Liu et al., 2025), their approach is non-interventional and doesn’t establish a causal link between data properties and cross-lingual transfer.

Our work connects and contributes to the body of research into spurious correlations (Geirhos et al., 2019; McCoy et al., 2019) which loom behind many surprising model failures and generalization challenges (Pearl & Mackenzie, 2018). Similar to recent work by Hermann & Lampinen (2020) and Lampinen et al. (2024), we investigate models’ inductive biases and aim to characterize how a feature’s complexity and predictivity influence its representation. However, we do so in a setup that imitates factual knowledge acquisition and transfer in LLMs, where feature complexity (i.e., low extractability) and predictivity (i.e., informativeness) *emerge naturally* from the standard training process. In addition, our results demonstrate that the language feature, while obsolete once the true feature is learned, persists in representations (App. 10.8). This has implications for fine-tuning when the model again may prefer to rely on an easier feature (Lovering et al., 2021).

Our results suggest two possibilities for improving cross-lingual recall: by obscuring differences between languages, or by balancing attribute frequencies across languages in the pretraining mixture. However language clearly *can* be a useful prior, e.g., in factual queries requiring a language- or culture-specific answer (though how much the model relies on it depends on how easily the language feature can be extracted).

While our two-language setup successfully isolates the pairwise mechanisms of cross-lingual transfer, real-world multilinguality is inherently combinatorial. Extending this analysis to multiple languages is an important future direction, though it requires modeling complex asymmetries, such as typological pivoting or interference across language families.

**Limitations** Our synthetic languages are defined solely as sets of templates, ignoring structural and lexical (dis)similarities between languages. Also, the attributes are spelled the same across our synthetic languages which is true for some attribute types (e.g., years or numbers) and some language pairs (e.g., Paris is spelled the same in several languages). While these are clear simplifications of LLM pretraining data, our key findings are independent of these design choices. Finally, our experiments focus exclusively on single-hop atomic recall (mapping a subject and relation to an attribute). They don’t indicate whether the unification mechanism extends to complex, compositional, or multi-hop factual reasoning, which represents a distinct and harder regime (Biran et al., 2024; Wang et al., 2024).

## 8 Conclusions

In this work we use a controlled setting to study why LMs often fail at cross-lingual knowledge transfer, hallucinating facts in one language that they know in another. We demonstrate how these failures are caused by models developing separate, language-specific representations for facts rather than unified, language-agnostic ones. Our key finding is that separation is driven by the *informativeness* and *extractability* of the language feature itself and happens very early in training. These results shed light on the role of tokenization, shared vocabulary, script and embedding alignment which have been observed previously but without explanation. Finally, we introduce a language unification score to quantify the representational similarity which is both strongly predictive of cross-lingual factual accuracy and useful for checkpoint selection.

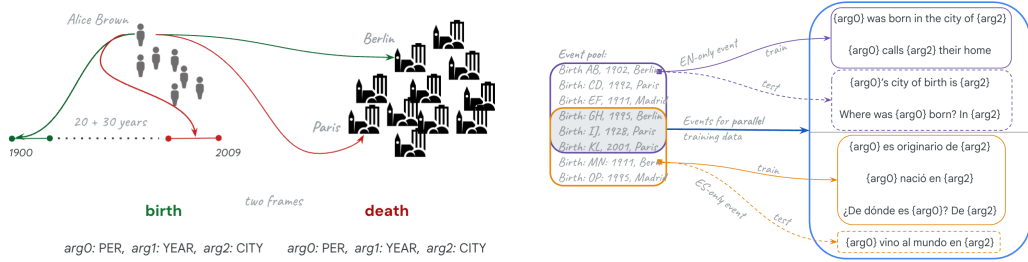
## References

- Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang. Language models’ factuality depends on the language of inquiry, 2025. URL <https://arxiv.org/abs/2502.17955>.
- Zeyuan Allen-Zhu. ICML 2024 Tutorial: Physics of Language Models, July 2024. Project page: <https://physics.allen-zhu.com/>.
- Stella Biderman, Mohammad Aflah Khan, Niloofar Miresghallah, Catherine Arnett, Fazl Barez, and Naomi Saphra. Position: Don’t just “fix it in post”: A science of ai must study training dynamics. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL <https://icml.cc/virtual/2026/poster/67142>.
- Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.781. URL <https://aclanthology.org/2024.emnlp-main.781/>.
- Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891. Curran Associates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/77c6ccacfd9962e2307fc64680fc5ace-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/77c6ccacfd9962e2307fc64680fc5ace-Paper-Conference.pdf).
- Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable influence and fact tracing for large language model pretraining. In *Proceedings of the 13th International Conference on Learning Representations*, 24–28 Apr 2025. URL <https://arxiv.org/abs/2410.17413>.
- 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. Association for Computational Linguistics, 2020a.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034. Association for Computational Linguistics, 2020b.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.
- Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bygh9j09KX>.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL <https://aclanthology.org/2023.emnlp-main.751/>.

- Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *International Conference for Machine Learning*, 2024. URL <https://arxiv.org/abs/2401.06102>.
- 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. Eclectic: a novel challenge set for evaluation of cross-lingual knowledge transfer, 2025. URL <https://arxiv.org/abs/2502.21228>.
- Katherine L. Hermann and Andrew K. Lampinen. What shapes feature representations? exploring datasets, architectures, and training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=HJeT3yrtDr>.
- Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/koh17a.html>.
- Andrew Kyle Lampinen, Stephanie C.Y. Chan, and Katherine Hermann. Learned feature representations are biased by complexity, learning order, position, and more. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=aY2nsgE97a>.
- 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, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.572. URL <https://aclanthology.org/2024.emnlp-main.572/>.
- Zheng Wei Lim, Alham Aji, and Trevor Cohn. Language-specific latent process hinders cross-lingual performance, 2025a. URL <https://arxiv.org/abs/2505.13141>.
- Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. Understanding cross-lingual inconsistency in large language models, 2025b. URL <https://arxiv.org/abs/2505.13141>.
- Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia K  rner, Ercong Nie, Barbara Plank, Fran  ois Yvon, and Hinrich Sch  tze. Tracing multilingual factual knowledge acquisition in pretraining, 2025. URL <https://arxiv.org/abs/2505.14824>.
- Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. Predicting inductive biases of pre-trained models. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=mNtmhaDkAr>.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Llu  s M  rquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL <https://aclanthology.org/P19-1334/>.

- Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. Exploring alignment in shared cross-lingual spaces. 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. 6326–6348, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.344. URL <https://aclanthology.org/2024.acl-long.344/>.
- Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 448–462. Association for Computational Linguistics, 2021a.
- Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2214–2231. Association for Computational Linguistics, 2021b.
- nostalgebraist. Interpreting gtp: The logit lens, 2020. URL <https://www.lesswrong.com/posts/AckRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>.
- Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024.
- 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, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.18. URL <https://aclanthology.org/2022.acl-long.18/>.
- Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Penguin Books, 2018.
- Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001. Association for Computational Linguistics, 2019.
- 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, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.658. URL <https://aclanthology.org/2023.emnlp-main.658/>.
- Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwaraknath Gnaneshwar, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=1hQKHHUsMx>.
- Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.
- Andrea Schioppa, Katja Filippova, Ivan Titov, and Polina Zablotskaia. Theoretical and practical perspectives on what influence functions do. In *Neural Information Processing Systems*, 2023. URL <https://arxiv.org/abs/2305.16971>.
- John Schulman. John schulman - reinforcement learning from human feedback: Progress and challenges. [https://www.youtube.com/watch?v=hhiLw5Q\\_UFg](https://www.youtube.com/watch?v=hhiLw5Q_UFg), 2023. Accessed: July 2025.

- Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*, 2025.
- Ammon Shurtz, Christian Richardson, and Stephen D. Richardson. When scripts diverge: Strengthening low-resource neural machine translation through phonetic cross-lingual transfer. In *Proceedings of the 5th Workshop on Multilingual Representation Learning*, pp. 336–346. Association for Computational Linguistics, 2025.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. *Advances in Neural Information Processing Systems*, 37:95238–95265, 2024.
- Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strotgen, and Hinrich Schutze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models, 2025. URL <https://arxiv.org/pdf/2504.04264>. To appear in ACL 2025.
- Andrea W Wen-Yi and David Mimno. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1124–1131, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.71. URL <https://aclanthology.org/2023.emnlp-main.71/>.
- Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, abs/2401.11817, 2024. URL <https://doi.org/10.48550/arXiv.2401.11817>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual 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. 10602–10617, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.coling-main.707/>.



(a) Knowledge Graph (KG) is populated with events of two types (b) Templates to express birth-city facts in EN and ES

Figure 9: **Birth** and **death** events are created for every entity by sampling from a set of years (disjoint) and cities (same pool). A dataset is comprised of monolingual (expressed in either **EN** or **ES**) and parallel (expressed in **both** languages) events. Arrows point from a particular event to the training templates (solid line) or in-language test templates (dashed). All verbalizations in the other language are part of the cross-lingual test set. To simplify, only birth templates with  $arg0$ ,  $arg2$  are shown.

## 9 Appendix A: Additional Details

### 9.1 Data generation procedure

For readability, Fig. 9 uses English and Spanish templates, but in our setup templates are constructed by randomly sampling tokens from a predefined vocabulary (see App. 9.3 for examples). No two templates share tokens, so no tokens are shared between languages (except for arguments). In all experiments we use **two languages**, mostly with five templates per language and fact type. See pseudo-code for the data generation process below; sample templates are shown in App. 9.3. Since cross-lingual events are seen in (at least) twice as many training examples as the events which are only expressed in a single language, we need to ensure that the total size of the training dataset is invariant to the fraction of parallel data so we can compare models fairly. To this end, we downsample verbalizations of monolingual events. In our experiments the attributes are sampled uniformly, which naturally results in a variability in attribute frequencies (e.g., there may be more people born in *Berlin* than in *Paris*). We have a set of 100 cities (joint between the birth and death events), and two disjoint sets of 20 & 30 years for birth and death years. We also create a custom tokenizer which tokenizes every word as a single token. We include pseudocode for the knowledge-graph generation and partitioning in Fig. 10.

### 9.2 Model & Training Details

The models for most of our experiments use the Gemma-2 architecture with models of approximately 2 million parameters. For most experiments, we use the following parameters:

| Hyperparameter                           | Value |
|------------------------------------------|-------|
| Hidden Size ( $D_h$ )                    | 128   |
| Intermediate Size ( $D_i$ )              | 512   |
| Head Dimension ( $D_{head}$ )            | 64    |
| Number of Hidden Layers ( $L$ )          | 6     |
| Number of Attention Heads ( $H_{attn}$ ) | 4     |
| Number of Key/Value Heads ( $H_{kv}$ )   | 1     |

We also conduct experiments with Pythia models and with Gemma models with both 0.1x the number of parameters and 10x the number of parameters, but scale did not end up being a substantial variable in this work. Models are trained for 100 epochs (although this is quite gratuitous, they typically converge in a small fraction of this time). We train models with a weight decay of 0.2, although we note that varying this from 0.0 to 0.4 did not change results very much. We train with a batch size of 128 and a cosine learning rate starting at  $3e-4$ .

---

```
def create_data_splits():
    events = []
    for name in names:
        birth = Event(name, random(cities), random(birth_days))
        death = Event(name, random(cities), random(lifespans) + birth.date)
        events.extend([birth, death])

    templates["lang0"]["birth", "time"] = []
    for _ in range(N_TEMPLATES):
        tokens = random(tokenizer.vocab(), size=(TEMPLATE_LEN,))
        tokens = randomly_insert("{subject}", tokens)
        tokens.append("{time}")
        templates["lang0"]["birth", "time"].append(tokens)

    crosslingual_events, other_events = random_split(events)

    crosslingual_train_data = cartesian_product(templates, crosslingual_events)
    other_train_data = {
        lang: cartesian_product(templates[lang], other_events[lang])
        for lang in LANGUAGES
    }

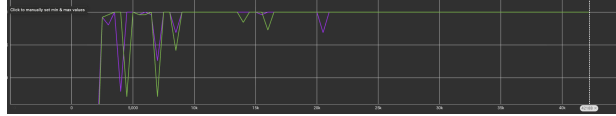
    other_train_data, same_lang_test = drop_equally_by_event(other_train_data)
    if is_too_big(other_train_data):
        other_train_data, extra_same_lang_test =
            drop_extra_train_data(other_train_data)
        same_lang_test += extra_same_lang_test

    cross_lang_test = {
        lang: cartesian_product(templates - templates[lang], other_events[lang])
        for lang in LANGUAGES
    }
    train_data = other_train_data + crosslingual_train_data
```

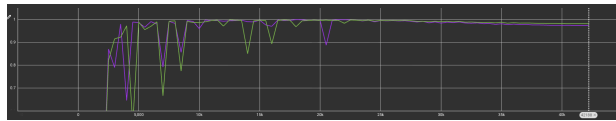
---

Figure 10: Pseudocode for data generation &amp; test set splitting

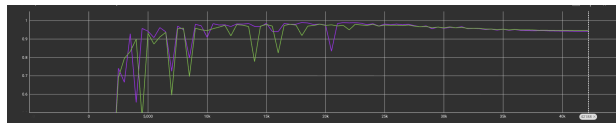
We include reference training plots in [11](#).



(a) Accuracy on the training set by step



(b) Accuracy on the same-lang test set by step



(c) Accuracy on the cross-lang test set by step

Figure 11: Model performance across various datasets. Subfigures (a), (b), and (c) display results for the training, same language test, and cross-lingual test sets, respectively.

### 9.3 Sample Dataset

Every time we create a dataset, we generate two new "languages" made of new templates. We show here all the templates for one of these datasets and a fragment of the resulting dataset.

| Language ID | Frame | Template                                                                  |
|-------------|-------|---------------------------------------------------------------------------|
| 0           | birth | {arg0} Ramoss recib 58 citizenship An 1965 verlassen {arg1}               |
| 0           | birth | 1922 cel 117 {arg0} top Barcel {arg2}                                     |
| 0           | birth | {arg0} died primer Rob Olivers Leuv ist 93 {arg1}                         |
| 0           | birth | {arg0} Johnstons mar Youn German New {arg2}                               |
| 0           | birth | ran hails {arg0} hap Hansens by {arg1}                                    |
| 0           | birth | deceas 107 read Montgom {arg0} Per {arg2}                                 |
| 0           | birth | p 65 Col 1967 {arg0} Hendersons Grah 57 {arg1}                            |
| 0           | birth | {arg0} Morgensterns born 2012 Leice welcomed {arg2}                       |
| 0           | birth | {arg0} 104 Chapmans Whose tuv Mur {arg1}                                  |
| 0           | birth | di {arg0} Wrights major Hawkinss Klagen {arg2}                            |
| 0           | death | {arg0} 108 encontró origins Frank 99 {arg1}                               |
| 0           | death | 64 1944 next or ta {arg0} c {arg2}                                        |
| 0           | death | {arg0} Blacks Andersons 2018 1960 Phoen {arg1}                            |
| 0           | death | {arg0} Lugan destination descans Coopers Podolskis Kelleys Duncans {arg2} |
| 0           | death | {arg0} Angel 16 Bishops 1951 Winnip Where 1927 {arg1}                     |
| 0           | death | Danielss Nant concerts Tur {arg0} 1933 nati {arg2}                        |
| 0           | death | music Rey Adam Perrys of {arg0} descanso fal {arg1}                       |
| 0           | death | Philadelph Thomass {arg0} Chavezs 19 Woodss {arg2}                        |
| 0           | death | Diazs Par 62 1969 {arg0} Miss Manche {arg1}                               |
| 0           | death | {arg0} 200 Johnstons 6 Who Meyers Harpers {arg2}                          |
| 1           | birth | {arg0} Montgom verstarb Diazs Foxs Blacks mundo 1949 {arg1}               |
| 1           | birth | {arg0} hat Aust meet V from Wh {arg2}                                     |
| 1           | birth | {arg0} 3 Watkinss Hend 21 Bankss Thomass Ott {arg1}                       |
| 1           | birth | Svet Morgans An {arg0} Lawrences Mart tied {arg2}                         |
| 1           | birth | Jamess Mont Br Ar raised {arg0} Daviss {arg1}                             |
| 1           | birth | {arg0} 85 Gardners Rices Wat 70 83 {arg2}                                 |
| 1           | birth | Sanderss {arg0} 81 65 48 fallecimiento Mey {arg1}                         |
| 1           | birth | Scotts 61 pass Ed {arg0} Rosss {arg2}                                     |
| 1           | birth | {arg0} studen Vasqu 40 Tuckers many erblickte 87 {arg1}                   |
| 1           | birth | {arg0} Hills 13 1971 Petersons 1951 encontró {arg2}                       |
| 1           | death | Torress es Camp married mund {arg0} 45 26 {arg1}                          |
| 1           | death | {arg0} seine Angeles 98 cam im {arg2}                                     |
| 1           | death | Hendersons {arg0} Sander tuv reader V raised {arg1}                       |
| 1           | death | {arg0} Po Jordans Dorn 76 Cal Whit Brow {arg2}                            |
| 1           | death | 2019 Longs 28 Knig listen {arg0} orig {arg1}                              |
| 1           | death | Gre Vanc {arg0} finali Ess Griff listening 42 {arg2}                      |
| 1           | death | chil k Tay Carters {arg0} Fiel recib When {arg1}                          |
| 1           | death | Geburts fallec dece {arg0} Lich Ortizs Parkers 35 {arg2}                  |
| 1           | death | {arg0} verlass Boy likes Co Svet Tou {arg1}                               |
| 1           | death | {arg0} Sc Martinezs nati Colemans 113 Mc {arg2}                           |

This is then an excerpt of six frames for this dataset. This dataset alone contains 2000 frames.

```
{
  "frame": {
    "core": "birth",
    "args": {"arg0": "PERSON", "arg1": "TIME", "arg2": "CITY"}
  },
  "arg_values": {"arg0": "Aaliyah Anderson", "arg1": "1906", "arg2": "Sheffield"}
}
{
  "frame": {
    "core": "death",
    "args": {"arg0": "PERSON", "arg1": "TIME", "arg2": "CITY"}
  },
  "arg_values": {"arg0": "Aaliyah Anderson", "arg1": "1990", "arg2": "Zurich"}
}
{
  "frame": {
    "core": "birth",
    "args": {"arg0": "PERSON", "arg1": "TIME", "arg2": "CITY"}
  },
  "arg_values": {"arg0": "Aubrey Griffin", "arg1": "1914", "arg2": "Portland"}
}
```

```

{
  "frame": {
    "core": "death",
    "args": {"arg0": "PERSON", "arg1": "TIME", "arg2": "CITY"}
  },
  "arg_values": {"arg0": "Aubrey Griffin", "arg1": "1996", "arg2": "Innsbruck"}
}
{
  "frame": {
    "core": "birth",
    "args": {"arg0": "PERSON", "arg1": "TIME", "arg2": "CITY"}
  },
  "arg_values": {"arg0": "Luke Dunn", "arg1": "1917", "arg2": "Edmonton"}
}
{
  "frame": {
    "core": "death",
    "args": {"arg0": "PERSON", "arg1": "TIME", "arg2": "CITY"}
  },
  "arg_values": {"arg0": "Luke Dunn", "arg1": "1997", "arg2": "Rostov"}
}

```

Finally, we create our datasets by expressing the events contained in these frames on a subset of templates. Here is a fragment of the resulting dataset:

| Language | Frame | Person           | Generated Text                                                      |
|----------|-------|------------------|---------------------------------------------------------------------|
| 0        | birth | Aaliyah Anderson | Aaliyah Anderson Ramoss recib 58 citizenship An 1965 verlassen 1906 |
| 0        | birth | Aaliyah Anderson | 1922 cel 117 Aaliyah Anderson top Barcel Sheffield                  |
| 1        | birth | Aaliyah Anderson | Aaliyah Anderson Montgom verstarb Diazs Foxs Blacks mundo 1949 1906 |
| 1        | birth | Aaliyah Anderson | Aaliyah Anderson hat Aust meet V from Wh Sheffield                  |
| 0        | birth | Aubrey Griffin   | Aubrey Griffin Ramoss recib 58 citizenship An 1965 verlassen 1914   |
| 0        | birth | Aubrey Griffin   | 1922 cel 117 Aubrey Griffin top Barcel Portland                     |
| 1        | birth | Luke Dunn        | Luke Dunn Montgom verstarb Diazs Foxs Blacks mundo 1949 1917        |
| 1        | birth | Luke Dunn        | Luke Dunn hat Aust meet V from Wh Edmonton                          |
| 0        | death | Aaliyah Anderson | Aaliyah Anderson 108 encontró origins Frank 99 1990                 |
| 0        | death | Aaliyah Anderson | 64 1944 next or ta Aaliyah Anderson c Zurich                        |
| 1        | death | Aaliyah Anderson | Torress es Camp married mund Aaliyah Anderson 45 26 1990            |
| 1        | death | Aaliyah Anderson | Aaliyah Anderson seine Angeles 98 cam im Zurich                     |
| 0        | death | Aubrey Griffin   | Aubrey Griffin 108 encontró origins Frank 99 1996                   |
| 0        | death | Aubrey Griffin   | 64 1944 next or ta Aubrey Griffin c Innsbruck                       |
| 1        | death | Luke Dunn        | Torress es Camp married mund Luke Dunn 45 26 1997                   |
| 1        | death | Luke Dunn        | Luke Dunn seine Angeles 98 cam im Rostov                            |

#### 9.4 Unification+ Score For Model Selection

At random initialization, the language unification score is typically around 1.0, because there is no more variation across-languages than there is within language. As a result of this, we utilize a slightly modified version of the language unification score when selecting the best checkpoint. For this, we multiply the language unification score with the same-lang test accuracy. Intuitively, this balances a model that fits the existing data well with one that generalizes well. This metric is listed as ‘unification-score+’ in the chart.

---

```

def compute_unification_scores(model, datapoints):
    total = 0
    for datapoints_by_fact in group_by_fact(datapoints):
        for idx in range(len(datapoints_by_fact)):
            datum = datapoints_by_fact[idx]
            same_lang = [
                d
                for i, d in enumerate(datapoints_by_fact)
                if d.language == datum.language and i != idx
            ]
            other_lang = [
                d
                for i, d in enumerate(datapoints_by_fact)
                if d.language != datum.language and i != idx
            ]
            total += mean_cosine_similarity(model, datum, other_lang) /
                    mean_cosine_similarity(model, datum, same_lang)

    return total / len(datapoints)

def mean_cosine_similarity(model, datum, other_data):
    return mean([
        cosine_similarity(
            concat_residual_latents_at_last_token(model, datum),
            concat_residual_latents_at_last_token(model, other_datum)
        )
        for other_datum in other_data
    ])

```

---

Figure 12: Pseudocode for calculation of language unification scores

## 10 Appendix B: Complementary Results

### 10.1 Gradient-based representations yield similar patterns

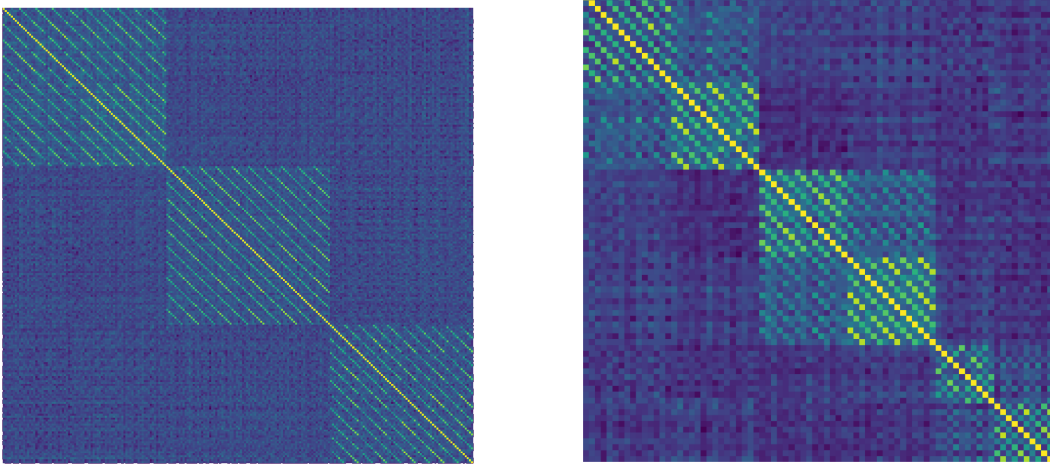


Figure 13: [Models with 30% and 4% cross-lingual events, final checkpoint] Also with **gradient-based** representations the language checkerboarding is visible in a model with poor cross-lingual generalization (right). Three most frequent (in the cross-lingual portion of the respective training data) birth-city attributes are picked to collect representations.

## 10.2 Layerwise results for activation language unification scores vs xLang Test Acc

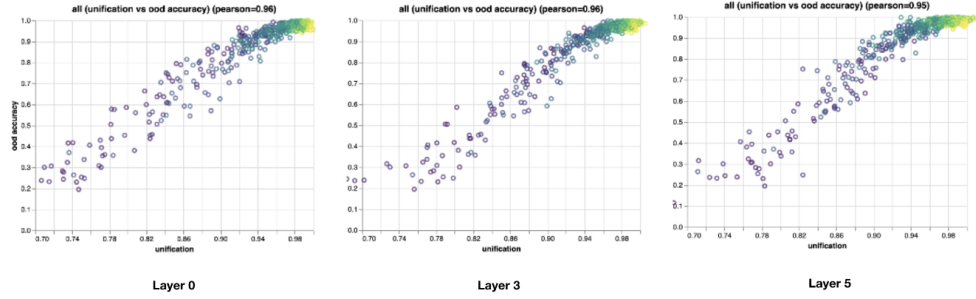


Figure 14: Layerwise results for activation language unification scores vs. xLang Test Acc.

## 10.3 Language checkerboarding is more obvious for lower layers

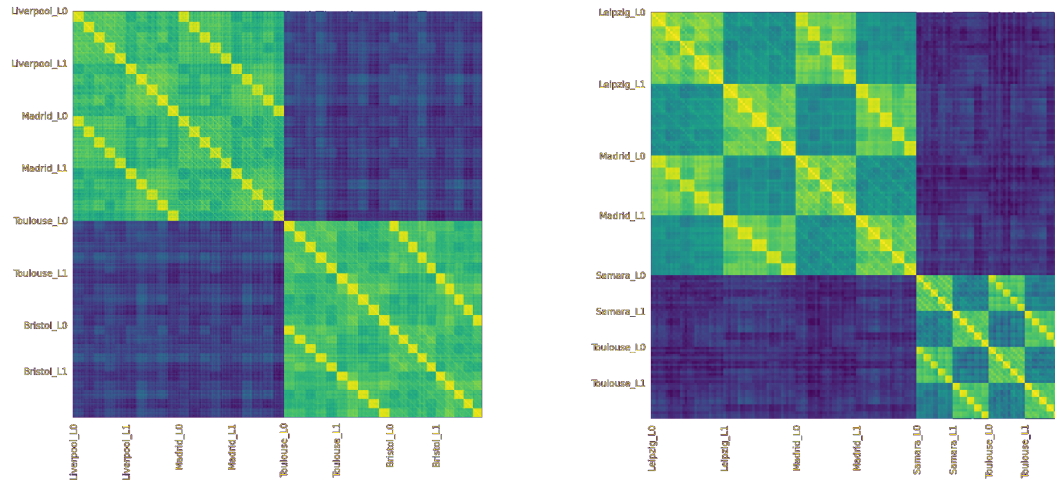


Figure 15: [Models with 30% and 8% cross-lingual events, layers 0-1, final checkpoint, activations-based representations] The presence of language checkerboarding (right) is particularly striking in the lowest model layers. For the poorly generalizing model, language identity is the dominant factor determining the similarity of representations for the same fact type. Two birth-city and two death-city attributes (most frequent in the cross-lingual portion of the respective training data) are picked to collect representations.

#### 10.4 Language checkerboarding is more obvious for imbalanced models

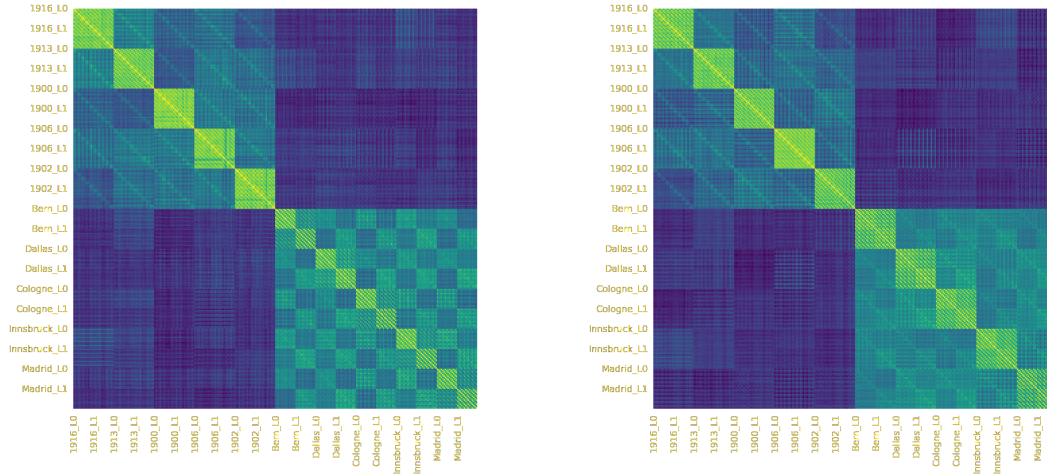


Figure 16: Imbalanced vs. balanced at checkpoint-40,000

#### 10.5 Language checkerboarding is more obvious for cased models

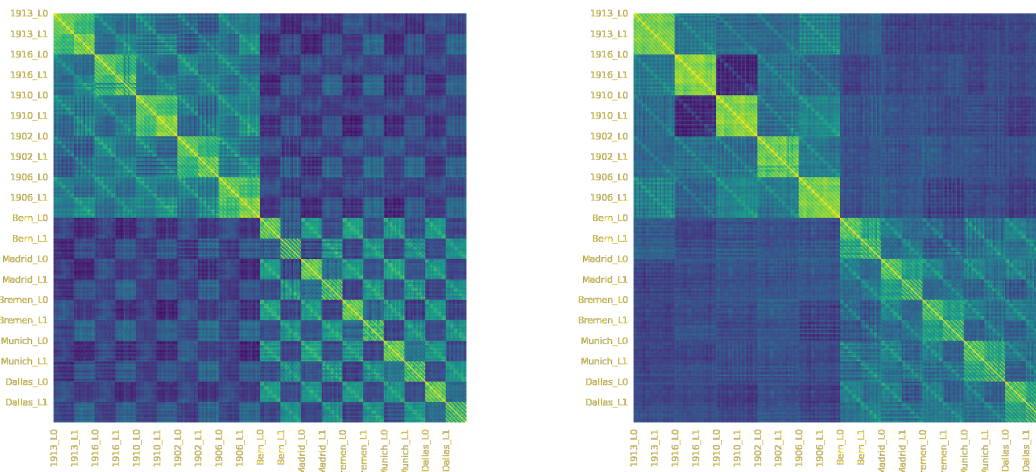


Figure 17: Cased vs all-lowercased (by language) at checkpoint-8,000

## 10.6 Adding templates improves cross-lingual generalization

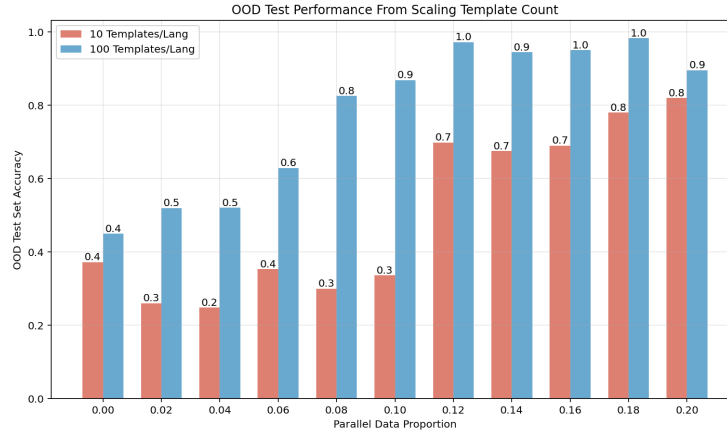


Figure 18: Increasing the number of templates substantially improves cross-lingual generalization, despite also increasing the difficulty of the test set.

## 10.7 Vocab Similarity vs Cross-Lingual Accuracy

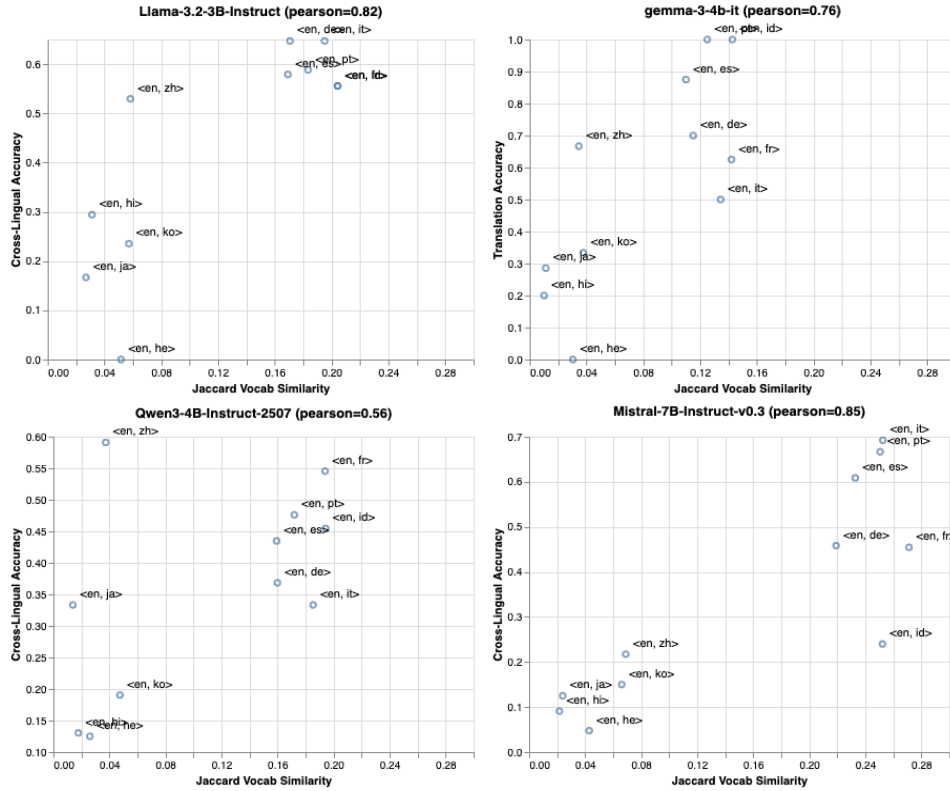


Figure 19: Cross-Lingual accuracy vs. vocabulary similarity.  $\langle x, y \rangle$  indicates going from source language  $x$  to target language  $y$ . For most models, the lower left includes languages like Hebrew that have little to no vocabulary overlap with English. On the other hand, languages such as Indonesian tend to have higher vocabulary overlap and higher cross-lingual accuracy. The average correlation coefficient is 0.69, ranging from 0.43 to 0.85 across different models.

Table 1: Percentage of variance explained in cross-lingual accuracy when considering vocabulary overlap between source and target languages.

| Source Lang. | de    | en    | es    | fr    | he    | hi    | id    | it    | ja   | ko    | pt    | zh   |
|--------------|-------|-------|-------|-------|-------|-------|-------|-------|------|-------|-------|------|
| # Datapoints | 90    | 265   | 95    | 45    | 80    | 210   | 90    | 120   | 140  | 65    | 95    | 85   |
| $R^2$ (%)    | 68.93 | 32.75 | 69.06 | 99.53 | 31.30 | 41.52 | 37.60 | 79.99 | 0.03 | 45.94 | 24.14 | 0.20 |

### 10.8 Fact ID vs language ID feature footprints.

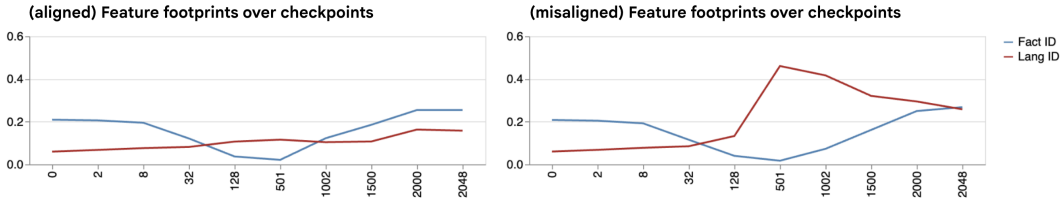


Figure 20: Representational variance explained by language (“Lang ID” - red) vs. fact (“Fact ID” - blue) across initial checkpoints. Language is uninformative for the model on the left (Balanced), whereas it is highly informative for the model on the right (Imbalanced). Both models are trained with 0% parallel data, thus language identity is highly extractable. However in the Imbalanced model, language identity is also highly informative. The language feature footprint grows quickly in the imbalanced model (before shrinking, although it still ends up larger in imbalanced than balanced models - see Fig. 8).