READ IN ENGLISH, ANSWER IN HINDI? EXPLORING DATA COMPOSITIONS FOR CROSS-LINGUAL KNOWLEDGE TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent evidence suggests that multilingual language models (LMs) can inexplicably transfer knowledge acquired in one language to other languages without explicit incentives for such behavior. Factors that govern such transfer aren't well understood, with past studies disagreeing on what constitutes knowledge transfer and what factors are important. Our work addresses this gap, and studies the role of patterns in multilingual pre-training data in unlocking cross-lingual knowledge transfer. We devise a controlled setup wherein we train language models from scratch on a synthetic corpus comprising biographies across multiple languages. Our experiments highlight that introducing parallel data improves transfer asymmetrically for dissimilar languages. Additionally, how parallel data is introduced also matters: smaller amounts of parallel data across many documents is more beneficial than higher parallelism across fewer documents. In addition to parallel data, the presence of code mixing across documents or complementary information across languages also improves transfer. We find that these findings also hold in real-world settings, where we continually pre-train off-the-shelf models on new knowledge, albeit transfer is moderate. The trends generalize similarly for a two-hop question-answering task, which requires a mix of cross lingual recall and reasoning. Overall, our findings offer data-curation strategies for pre-training that enable knowledge transfer in multilingual language models.[1]

1 INTRODUCTION

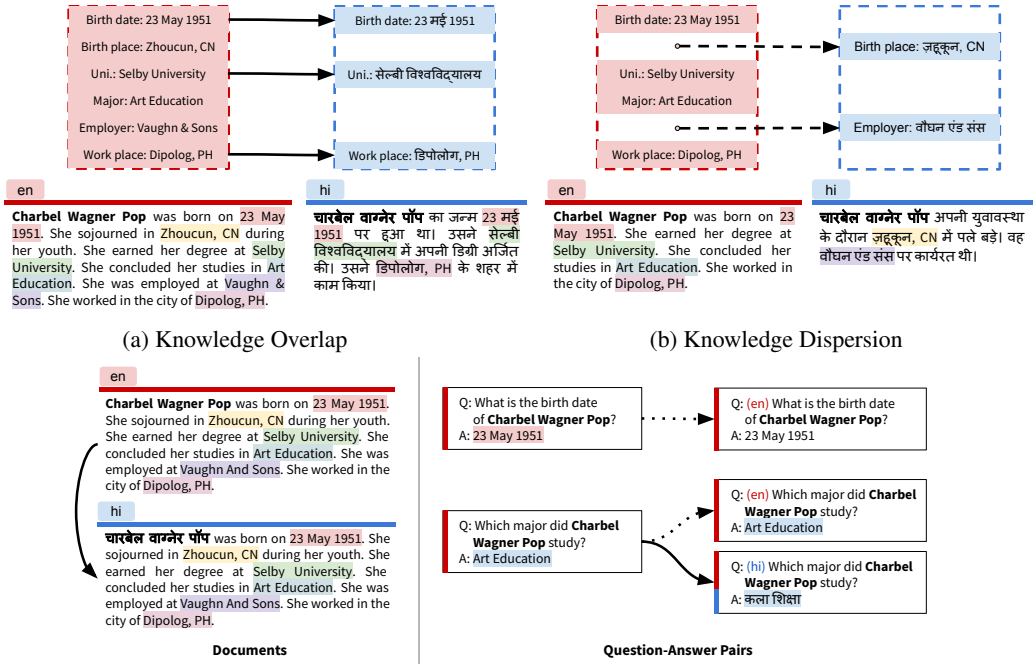
Multilingual language models acquire world knowledge by pre-training on massive text corpora (Grattafiori et al., 2024; Team et al., 2025, *inter alia*). However, the paucity of high-quality data limits equitable knowledge acquisition in languages that are underrepresented. For example, in the large-scale multilingual corpus mC4 (Xue et al., 2021), the number of pre-training tokens in Russian, the second-highest resource language after English, is 4 times lower than English. The representation of other languages is even smaller, typically orders of magnitude lower.

Despite this lack of data, recent studies find that multilingual models can occasionally perform *cross-lingual knowledge transfer* (XLT): recall and apply knowledge learnt in one language across other languages, without explicitly training on facts in the language being queried (Goldman et al., 2025). The emergence of this transfer is surprising, as current recipes for model pre-training do not explicitly optimize for this behavior (Grattafiori et al., 2024). In the absence of explicit incentives, it is unclear why we observe this transfer of knowledge, and what factors can promote such transfer. Since many languages may never have the extent of resources that English enjoys, understanding and amplifying factors underlying cross-lingual transfer can help us build better multilingual models.

Insights from prior work studying cross-lingual transfer do not fully explain its emergence. This is as prior work is divided on what constitutes knowledge transfer. For instance, some studies describe knowledge transfer to mean transfer of the knowledge required to perform a given task (Zhuang & Sun, 2025; Kallini et al., 2025), as opposed to transfer of the underlying factual knowledge (Blum et al., 2025; Liu et al., 2025). The specific task employed to study transfer also differs – some studies

[1]We will release all data and code to facilitate reproducibility of our findings upon publication.

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.



(a) Knowledge Overlap (b) Knowledge Dispersion

(c) Language Mixing, across documents or QA pairs

Figure 1: Overview of data compositions that we study. Each composition is applied over a document in a source language (en), resulting in the addition of biographies or QA pairs in other languages (such as hi). Using recipes that introduce these compositions, we study their role on transfer.

operationalize transfer via translation (He et al., 2025; Shen et al., 2025) or instruction-following (Chirkova & Nikoulina, 2024) – which may not explain knowledge transfer (Chua et al., 2024). Other studies focusing specifically on factual knowledge involve synthetic languages (Blum et al., 2025), leading to insights which may not necessarily transfer to natural data. These differences make it difficult to explain knowledge transfer based on prior work.

In this work, we study how the composition of pre-training data affects the emergence of XLT in multilingual models. To test the effects of varying compositions, we first design a controlled setup for evaluating transfer. This setup, inspired from the study by Allen-Zhu & Li (2025a), comprises a synthetic multilingual dataset of fictitious biographies, on which we train small language models *from scratch*. This allows us to eliminate test-set contamination and cleanly attribute the role of data compositions towards transfer. We then validate our findings using a real-world setup, where we continually pre-train an off-the-shelf model over a corpus of recent articles curated from Wikipedia. Such a setting mimics real-world applications of updating models for new knowledge (Yıldız et al., 2025; Li et al., 2025). With this setup, we validate the utility of different data compositions.

We study the role of compositions by devising various data curation recipes. Each recipe creates a corpus with a particular composition of data (see Figure 1). We explore 3 compositions: *overlap*, *dispersion*, and *language mixing*. *Knowledge overlap* models parallel data present in multiple languages, and is introduced by translating some of the facts (Fig. 1a). *Knowledge Dispersion* represents the presence of complementary information across languages (Wang et al., 2025b), leading to unique facts per language in each document (Fig. 1b). Lastly, *Language Mixing* models code-mixing in either documents or task data (Fig. 1c). To evaluate transfer, we probe the model for recall and factual reasoning, in languages different from the one where the knowledge was acquired. The model's answering accuracy represents the extent of transfer from the underlying recipe.

From our experiments, we summarize the following key findings:

- Overlap of facts across languages (i.e., parallel data) significantly improves knowledge transfer. Transfer improves by 32% in the controlled setting, and 9% in the real-world setting (§5.1).
- Cross lingual transfer due to overlap is asymmetric for dissimilar languages such as English and Hindi. That is, knowledge acquired in Hindi can be better recalled in English (39%) compared

to the opposite direction (10%). Such asymmetry remains even when models are trained with balanced data using a tokenizer that has comparable fertility across languages. (See §5.2).

- The way overlap is introduced also influences transfer: smaller amounts of overlap across many documents is more effective than higher overlap across fewer documents, yielding a recall of 41% and 30% respectively. This effect is more pronounced for dissimilar languages (§5.2).
- The use of language mixing – code-mixing within a document or a QA pair – improves the balance of transfer between languages, complementing overlap. However, this effect is less pronounced in the real-world setup (§5.3).
- For the first time, we establish that the presence of knowledge dispersion – complementary information across languages – also facilitates transfer, improving recall to 24% compared to the baseline. Consistently in the real-world setting, dispersion yields a gain of 5% above the baseline. This transfer is due to an increase in language diversity across training batches. (§5.4)
- Combinations of data recipes complement the effects of each other, further improving transfer. We find a combination of overlap and mixing yields the highest improvements, increasing recall to 53%. In the real-world setup, this recipe yields a gain of 11% (§5.5).
- While the major focus of our work is to understand knowledge transfer, we observe that factors that facilitate XLT also improve performance on two-hop question answering, which requires some amount of reasoning in addition to cross-lingual knowledge recall (§5.1).

These findings provide new insights into the effectiveness of different data curation recipes for cross knowledge transfer. We recommend that practitioners pre-training multilingual models (a) spend their curation efforts to obtain broad-coverage parallel data, by including as many documents as possible instead of increasing the amount of overlap across fewer documents, (b) leverage language mixed data during pre-training to improve the balance of transfer, and (c) use multilingual data that complements information in other languages to improve transfer.

2 Related Work

Explaining Knowledge Transfer in Language Models. While cross-lingual transfer is a well-studied domain, insights from past work cannot fully explain the emergence of XLT in current models. Many of these works disagree on *what constitutes transfer*. For instance, the study by He et al. (2025) measures transfer based on a word-level translation task. Other works measure transfer using zero-shot improvements to language modeling (De Souza et al., 2024), translation (Yuan et al., 2024), instruction following (Chirkova & Nikoulina, 2024), classification (Shen et al., 2025; Patil et al., 2022) or question answering (Kallini et al., 2025; Wang et al., 2025a). However, factors influencing cross-lingual transfer differ across tasks (Blaschke et al., 2025; Ahuja et al., 2022). Even for knowledge transfer, the precise definition employed may refer to transfer of the knowledge for the task (Zhuang & Sun, 2025; Kallini et al., 2025; Deshpande et al., 2022) instead of factual knowledge (Blum et al., 2025; Liu et al., 2025), disagreeing with more recent definitions (Goldman et al., 2025). Across studies that focus on factual knowledge, modeling assumptions limit their applicability. For instance, the study by Blum et al. (2025) explores the effects of parallel data and vocabulary overlap on knowledge transfer. However, their analysis only uses synthetic languages, insights from which may not extend to natural language data. Other studies leveraging synthetic languages face similar shortcomings (Deshpande et al., 2022; Schäfer et al., 2024). To address these gaps, we study the role of data compositions towards transfer using a systematic approach with natural language data.

Evaluating Contamination-Free Knowledge Transfer. Successful knowledge transfer requires that a model is able to apply its parametric knowledge in a language different from where it is learnt (Goldman et al., 2025). Unfortunately, the training data for current models is often unknown (Team et al., 2025; Grattafiori et al., 2024), making identification of sources difficult. Even in monolingual corpora, unaccounted language contamination may contribute to transfer (Blevins & Zettlemoyer, 2022). Further, even when the training data is known, attributing knowledge sources pose a challenge. This is as knowledge learnt by a model may be acquired directly or indirectly from one or more sources (Yang et al., 2024). Past work studying cross-lingual transfer accounts for these challenges by controlling the training data (Kallini et al., 2025; Deshpande et al., 2022, *inter alia*). However, explicit efforts to ensure knowledge isolation are limited. In our work, we systematically restrict a subset of the knowledge to a single language, allowing us to quantify true knowledge transfer.

3 METHODOLOGY

In this section, we describe the approach we use to study the role of data compositions towards *cross-lingual transfer* (XLT) of knowledge. Our experiments train a decoder-only language model on a multilingual dataset of documents and associated question-answer (QA) data. This dataset is constructed using one of several data recipes, each modeling a particular composition of natural multilingual data. These recipes alter the source dataset to introduce the desired composition. By contrasting different recipes, we study the effects of dataset compositions on XLT.

We contrast recipes in two different setups. We first devise a controlled setup to identify effective data recipes, using a corpus of synthetic biographies. Biographies are a natural choice, as they can be modeled as a tuple of attributes while incorporating the nuances of natural language. These properties allow a systematic evaluation of factual recall using a corpus built with natural sequences. A similar approach has been used to study knowledge acquisition in prior work (Allen-Zhu & Li, 2025a). We then validate our findings using a real-world setup that uses a corpus of Wikipedia articles.

Notation. Let $\mathcal{D}$ denote a dataset of M documents. These M documents are synthetic biographies in the controlled setup, and Wikipedia articles in the real-world setup. Each document $d \in \mathcal{D}$ refers to a unique person or article as per the setup, and comprises $K(d)$ segments. A segment describes one or more attributes about the entity. We use N to refer to the number of unique documents and $\mathcal{L}$ to refer to the set of languages in $\mathcal{D}$. For a document $d \in \mathcal{D}$, we use $l(d)$ to refer to its language.

To test knowledge acquisition, we use QA pairs corresponding to different attributes. We use the following definition for XLT, following recent work (Goldman et al., 2025; Gao et al., 2024):

Definition 1 (**Cross-Lingual Transfer (XLT)**). Successful cross-lingual transfer from a language l to a language l' requires a model to answer a question in l' using knowledge acquired *only* in l.

For every document $d \in \mathcal{D}$, we curate a set of QA pairs $\mathcal{Q}_\mathcal{D}$ related to attributes of d, in the same language $l(d)$. We sample half of the data from $\mathcal{Q}_\mathcal{D}$ for training, referred as $\mathcal{Q}_D^{(\text{tr})}$. The other half, $\mathcal{Q}_D \setminus \mathcal{Q}_\mathcal{D}^{(\text{tr})}$, is held-out for monitoring and evaluation, and is referred as $\mathcal{Q}_\mathcal{D}^{(\text{ml})}$. For evaluating XLT, QA pairs in $\mathcal{Q}_\mathcal{D}$ are translated to other languages from $\mathcal{L}$, resulting in the test set $\mathcal{Q}_\mathcal{D}^{(\text{te})}$.

Let $\mathcal{M}$ be a model to evaluate, and $\mathcal{M}(q)$ denote the model's response to a question q. We have:

Definition 2 (**Cross-Lingual Recall, XLR**). The cross-lingual recall is the QA-accuracy on $\mathcal{Q}_\mathcal{D}^{(\text{te})}$:

$$\text{XLR}(\mathcal{M}, \mathcal{D}) = \frac{1}{|\mathcal{Q}_\mathcal{D}^{(\text{te})}|} \sum_{(q,a) \in \mathcal{Q}_\mathcal{D}^{(\text{te})}} \text{Eq}(\mathcal{M}(q), a) \tag{1}$$

where the function Eq compares the similarity of the model's response with the expected answer. Answering questions in $\mathcal{Q}_\mathcal{D}^{(\text{te})}$ requires a model to recall the knowledge learnt from $\mathcal{D}$ in a different language, in line with our definition of XLT.

We also evaluate the model's knowledge acquisition capabilities for each language in $\mathcal{L}$, using held-out questions from $\mathcal{Q}_\mathcal{D}^{(\text{ml})}$. This captures the model's *multilingual recall*. We have:

Definition 3 (**Multilingual Recall, MLR**). Multilingual recall is the QA-accuracy on $\mathcal{Q}_\mathcal{D}^{(\text{ml})}$:

$$\text{MLR}(\mathcal{M}, \mathcal{D}) = \frac{1}{|\mathcal{Q}_\mathcal{D}^{(\text{ml})}|} \sum_{(q,a) \in \mathcal{Q}_\mathcal{D}^{(\text{ml})}} \text{Eq}(\mathcal{M}(q), a) \tag{2}$$

We focus on three key compositions: *knowledge overlap*, *knowledge dispersion*, and *language mixing*. For comparison, we use a control recipe without overlap, dispersion or mixing as the baseline.

Baseline ($\mathcal{D}_{\text{base}}$). The baseline data recipe constructs a document corpus $\mathcal{D}_{\text{base}}$, where every entity's document is present in *only one* of $\mathcal{L}$ languages, resulting in an unparallel corpus. This implies that for $\mathcal{D}_{\text{base}}$, $M = N$, i.e., every document refers to a unique person or Wikipedia article. We treat this recipe as a baseline to compare with other recipes that aim to incentivize XLT.

It is important to note that data recipes are applied over a fraction, f, of documents in the baseline corpus, denoting the recipe's *coverage*. For $f \in (0, 1]$, the recipe will be applied only over $f \cdot M$ documents. The remaining $(1 - f) \cdot M$ documents are left unmodified, as in $\mathcal{D}_{\text{base}}$.

Knowledge Overlap ($\mathcal{D}_{\text{overlap}}$). We induce overlap by translating a fraction, p, of segments per document in $\mathcal{D}_{\text{base}}$ into all other languages in $\mathcal{L}$. As shown in Fig. 1a, we introduce overlap by translating half of the attributes ($p = 1/2$) of CHARLES WAGNER POP, such as birthdate, to Hindi. These translations create a new biography with 3 attributes, conveying the same but lesser information as English. These new documents are merged with $\mathcal{D}_{\text{base}}$, constituting $\mathcal{D}_{\text{overlap}}$. While QA pairs in $\mathcal{Q}^{(\text{tr})}$ corresponding to translated segments in the original documents are also translated into other languages, we remove the translated QA pairs from the test set, $\mathcal{Q}^{(\text{te})}$, to avoid any contamination.

Knowledge Dispersion ($\mathcal{D}_{\text{disp}}$). Dispersion models the spread of information across different languages. Such complementary arrangement of information is a known characteristic of Wikipedia articles (Wang et al., 2025b), and we therefore test whether it contributes to XLT. We introduce dispersion by moving a fraction, q, of segments per document in $\mathcal{D}_{\text{base}}$ to another language via translation. This process is repeated independently for every other language in $\mathcal{L}$, resulting in documents that *do not* share information. Fig. 1b demonstrates dispersion for the biography of CHARBEL WAGNER POP. Assuming English as the source and $q = 1/3$, two attributes, birthplace and employer, are "dispersed" from English to Hindi. The resulting documents form the corpus $\mathcal{D}_{\text{disp}}$.

Language Mixing ($\mathcal{D}_{\text{mix}}$). This recipe models the presence of code-mixing in multilingual corpora (Wang et al., 2025a). We introduce mixing either at the document or the QA level. When mixing across documents, for every document in $\mathcal{D}_{\text{base}}$, we create code-mixed documents in every other language in $\mathcal{L}$. In these documents, the name of the person or article is replaced with its translation in the other language. Fig. 1c (left) demonstrates language mixing for CHARBEL WAGNER POP. These new documents are added to the base corpus, creating $\mathcal{D}_{\text{mix,BIO}}$.

We introduce QA mixing by identifying a fraction, r, of segments per document in $\mathcal{D}_{\text{base}}$. For questions about these segments in the train set, $\mathcal{Q}^{(\text{tr})}$, we create code-mixed questions by translating the answer to all other languages in $\mathcal{L}$. Fig. 1c (right) demonstrates QA mixing for CHARBEL WAGNER POP. With $r = 1/2$, one of the questions in English is mixed. The new questions are merged with the train set, forming $\mathcal{D}_{\text{mix,QA}}$. We also add a language marker to each question to denote the expected answer language. To avoid any contamination, questions associated with the selected segments used for mixing are removed from the test set, $\mathcal{Q}^{(\text{te})}$.

Combinations. We also consider combinations of recipes, defined by applying two or more recipes collectively over the base corpus. For example, using $p = 1/2$ and $q = 1/6$, we obtain a corpus $\mathcal{D}$ where 50% of segments are overlapped across all languages and 16% of segments are dispersed equally. These recipe combinations correspond to the presence of multiple compositions in a dataset.

Translation Costs. The translation costs of applying our recipes are discussed in Appendix §C.

4 EXPERIMENTAL SETUP

We study the effectiveness of data recipes using two setups. We first devise a controlled setup to identify effective data compositions, followed by a real-world setup to validate our findings.

4.1 CONTROLLED SETUP

We perform controlled experiments using a corpus comprising synthetically-generated biographies and associated QA pairs in various languages. Over these datasets, we train small models *from scratch*. Effective recipes are identified through the knowledge recall of the trained model, highlighting compositions that positively benefit XLT.

We choose $\mathcal{L} = \{\text{English, French, Hindi}\}$ for our experiments. English and French share subwords, script and word order with each other, but have no overlapping linguistic traits with Hindi. Therefore, our choice of $\mathcal{L}$ allows our study to focus on findings that are agnostic to linguistic features. We synthetically curate biographies for $N = 10^5$ individuals. The total number of biographies, M, depends on the data recipe being used. All our recipes ensure that the training data is balanced across languages. While this does not match real-world language distributions, we use such uniformity to control for imbalance effects (Schäfer et al., 2024). Further details are in Appendix §A.1.

We train LLaMA-like decoder-only models over our curated corpora from scratch. Each model has 135M parameters. We also train our own tokenizer, to account for the nature of our data and to ensure

comparable token fertility for all languages. All biographies from the corpus are randomly shuffled and collated into sequences of 512 tokens. As prescribed by Allen-Zhu & Li (2025a), each sequence contains either only biographies or only QA, to prevent shortcut learning. We use a batch size of 96 during training. All training hyperparameters are determined once apriori using a random sweep. Across experiments, we only change parameters related to the data recipe. We train for multiple epochs, until convergence is reached, i.e., until the model's completion accuracy for biographies in the training corpus reaches $\simeq 100\%$. We train 3 models over the same dataset, initialized via different seeds, and report the average performance of the best checkpoint across each run (based on a held-out development set). For a fair comparison, we report evaluations over the intersection of $\mathcal{Q}^{(\text{te})}$ across recipes being compared, further stratified by language. During evaluation, we compare whether the model's response *exactly matches* the expected answer. Further details are in Appendix §B.1.

4.2 Real-World Experiments

The simplicity and scale of the controlled setup differ from real-world settings, where models are trained on much larger and potentially noisy corpora (Grattafiori et al., 2024). To account for these differences, we consider the task of continual pre-training to incorporate new knowledge (Li et al., 2025). The corpus for this task comprises natural data sourced from Wikipedia. With this setup, we verify whether our findings hold in realistic settings.

We experiment with continually pre-training models on a corpus of $M = 70$K documents from Wikipedia. Unlike the controlled setting, models may have trained on Wikipedia data. To reduce contamination, we only consider Wikipedia articles created after January 2024. Data recipes are applied treating passages within a document as segments. For every passage within a document, we generate up to 4 questions for evaluating recall. These questions are generated via an LLM using an approach inspired by Calderon et al. (2026). We perform several steps of validation to ensure these questions are precise, unambiguous and grounded in context. Of 300 manually validated QA pairs, 92% satisfy the required constraints. We also filter questions answerable by the model before training to reduce contamination. Additional details related to the construction are in Appendix §A.2.

We use $\mathcal{L} = \{\text{English, Kannada, Hindi}\}$ for the real-world experiments. We choose Kannada over French, to test whether our recipes also benefit low-resource languages. Note that in the controlled setup, all languages use equal data proportions and thus no language is low-resource. To facilitate learning (Allen-Zhu & Li, 2025b), we include 4 paraphrases of each article for training.

Our experiments continually pre-train Gemma 3 4B (Team et al., 2025) models. As in the controlled setup, we collate QA and documents separately. We train models for 3 epochs, with a batch size of 64, context length of $16K$ and a learning rate of 10^{-4}. Across experiments, we only change the data recipe. The training data comprises anywhere from 1.6B to 2.5B tokens, depending on the data recipe (details in §A.3). Results are reported from the best epoch, based on a development set. We evaluate responses against expected answers using an LLM judge, following Wei et al. (2024). We use Gemma 4 MoE (Team et al., 2026) and validate this choice against several other judges, finding near-perfect agreement. Our results are thus robust to the choice of judge (details in Appendix §B.2).

Additionally, we evaluate these models on a harder task of factual reasoning, using two-hop questions. These questions are derived by merging pairs of single-hop evaluation questions. Such questions require reasoning over recalled knowledge to deduce the answer, and help us assess whether the trained models generalize to tasks beyond simple recall.

5 Results

We present several interesting findings in this section, highlighting the effects of data compositions on cross-lingual knowledge transfer (XLT). We compare against the dataset without any composition ($\mathcal{D}_{\text{base}}$) as the baseline across experiments. We refer to languages by their ISO-639-1 codes.

5.1 Effects of Data Compositions

Figure 2 summarizes the effects of various data compositions across the two setups. Compared to the baseline, all data compositions we study positively benefit XLT. Effects in the controlled setup are more prominent, where recipes like overlap improve XLT from only 9.2% to 41.4% on

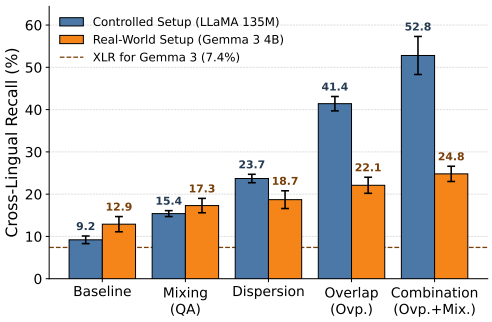


Figure 2: XLR for data recipes in the controlled and real-world setups. **Across both setups, all data compositions improve transfer.** Errors denote 2σ deviations over 3 training runs (controlled) or 200 bootstrap samples (otherwise).

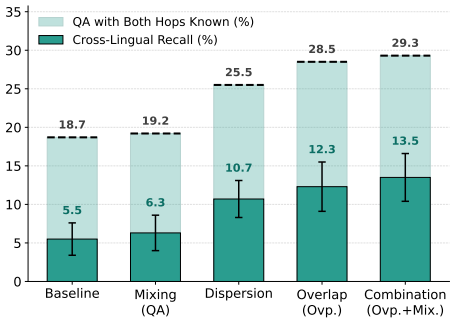


Figure 3: Performance over two-hop questions for data recipes in the real-world setting. **Recipes that improve single-hop recall also improve multi-hop performance**. Errors denote 2σ deviations over 200 bootstrap samples.

average. Overlap yields the highest improvements, followed by dispersion (23.7%) and language mixing across QA (15.4%). These trends extend to the real-world setting, with consistent but less pronounced relative gains in XLR. We speculate this is because language-wide disparities after pre-training are hard to undo through fine-tuning. Combining recipes yields the largest gains across all languages. These improvements do not affect general multilingual capabilities (details in §D.3).

Improvements from recipes generalize beyond simple recall. We further evaluate models from the real-world setup over two-hop questions, which require reasoning over learnt knowledge to answer. Fig. 3 shows the XLR over such questions. Remarkably, recipes that improve single-hop recall also perform better on multi-hop questions. This may be attributed to an increase in the fraction of two-hop questions where both hops are known, suggesting that improvements from recipes extend beyond simple recall. More details are in Appendix §D.4.

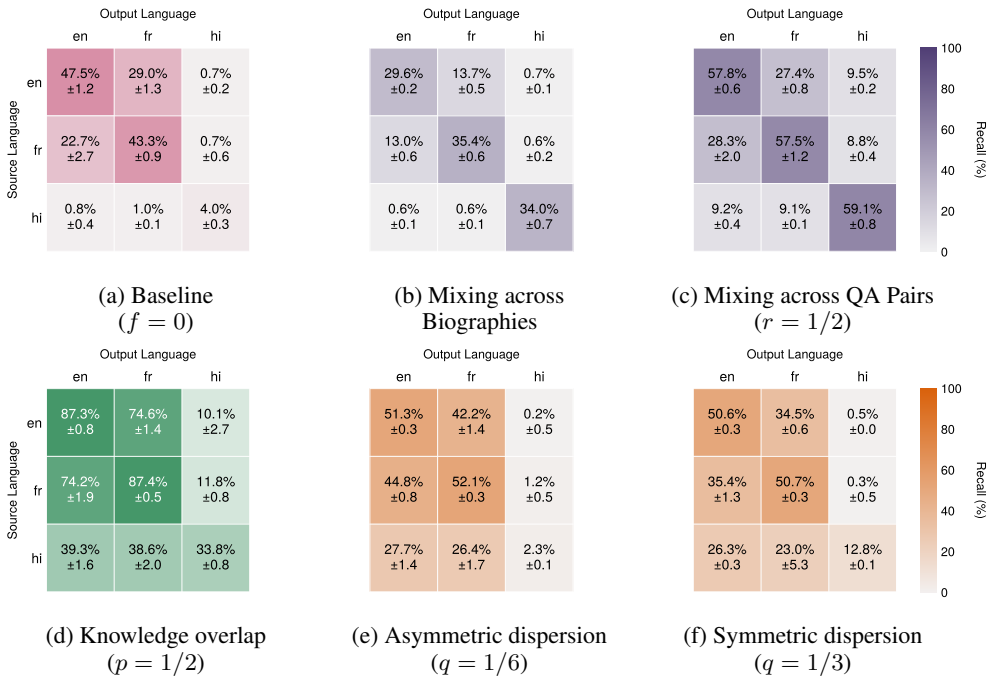
5.2 Effects of Knowledge Overlap

Introducing overlap improves knowledge transfer asymmetrically for dissimilar languages. In Fig. 4a & 4d, we compare models trained on $\mathcal{D}_{\text{base}}$ to ones trained on $\mathcal{D}_{\text{overlap}}$. With $\mathcal{D}_{\text{base}}$, transfer between `en` and `fr` is far from ideal, even though the two languages share many linguistic traits. With $\mathcal{D}_{\text{overlap}}$, knowledge transfer increases between `en` and `fr`, which shows that parallel data is useful for improving transfer *even between similar languages*. We observe a smaller improvement in cross-lingual recall (XLR) for pairs involving `hi`.

Interestingly, the transfer in directions `hi` $\rightarrow$ {`en, fr`} is considerably higher than in the opposite direction ({`en, fr`} $\rightarrow$ `hi`). In addition, the MLR for `hi` remains low, suggesting poor knowledge acquisition. While we do not offer a definitive explanation for why this is the case, we rule out several plausible causes. We confirm that asymmetry is not due to data imbalance (equal documents per language) or tokenizer bias (we re-train our tokenizer to have comparable fertility across languages). Although some of the asymmetry may be attributed to scripts of languages, asymmetry persists when training only on `fr` and `hi`. We discuss these further in Appendix §D.5.

Increasing overlap yields diminishing returns for cross-lingual transfer beyond a point. In Fig. 5, we observe an absence of XLT across the baseline, observing transfer only when some overlap is introduced (over the minimum value of 17% allowed by our controlled setup). This suggests a minimum amount of overlap may be necessary for transfer. Further increasing overlap up to 50% yields a significant boost in XLR, but beyond that, the gains are relatively minor. Based on this result, we recommend that practitioners need not invest resources to achieve 100% parallelism, as even moderate extents of overlap result in comparable XLT at significantly lower translation costs.

For the same amount of overlap, higher coverage is more beneficial than higher extent. Introducing overlap in a data corpus incurs a curation cost (translating, scraping, etc.), which we assume to be proportional to the coverage and extent of overlap. Model developers that have a fixed budget can spend it along two axes that both increase overlap: (i) They may achieve higher *extent* by translating



(a) Baseline ($f = 0$)

(b) Mixing across Biographies

(c) Mixing across QA Pairs ($r = 1/2$)

(d) Knowledge overlap ($p = 1/2$)

(e) Asymmetric dispersion ($q = 1/6$)

(f) Symmetric dispersion ($q = 1/3$)

Figure 4: Breakdown of recall across data recipes in the control setting. Diagonal cells denote MLR, while non-diagonal cells represent XLR. Transfer via overlap is asymmetric for dissimilar languages. Mixing improves MLR and balances transfer, and is better with QA. Dispersion also improves XLR, with higher dispersion being effective for MLR. Errors denote 2σ deviations from 3 training runs.

more parts per document or (ii) achieve greater *coverage* by translating fewer parts across more documents. In Figure 6, we compare three strategies, all requiring the same translation cost. We find that increasing coverage increases the *average* XLR over all language pairs. In particular, the strategy with the highest coverage yields the highest transfer for dissimilar pairs of languages (i.e., pairs involving `hi`). Based on these findings, we recommend that practitioners operating under a fixed budget are better off spending it to increase coverage for overlap, keeping the extent minimal.

5.3 Effects of Language Mixing

Language mixing improves multilingual recall. From Fig. 4a-4c, we observe that mixing significantly improves multilingual recall in `hi`. While transfer reduces between `en` and `fr`, mixing across QA modestly improves transfer between dissimilar languages (e.g., {`en, fr`} $\rightarrow$ `hi`). Further, this transfer is less asymmetric than recipes like $\mathcal{D}_{\text{overlap}}$. While not very useful in isolation, these properties are appealing as mixing can complement other recipes. As discussed in §5.5, combining $\mathcal{D}_{\text{mix,QA}}$ with other recipes improves MLR and XLR beyond that of the original recipe(s).

Language mixing across QA pairs is more beneficial than mixing across biographies. From Figure 4b and 4c, we observe that $\mathcal{D}_{\text{mix,QA}}$ yields a higher XLR on average across language pairs, while maintaining the symmetry of transfer in pairs of dissimilar languages.

Increasing the number of mixed questions consistently improves cross-lingual transfer. We also compare the change in XLR as we increase the number of mixed questions in Fig. 5. Unlike $\mathcal{D}_{\text{overlap}}$, increasing mixing (i.e., r) continues to reliably improve XLT, particularly improving XLR for `hi`.

5.4 Effects of Knowledge Dispersion

Knowledge dispersion also improves knowledge transfer. Surprisingly, dispersion also improves XLR as well as MLR on average, as observed in Fig. 2 and 4. In particular, MLR improves with the amount of facts dispersed (Fig. 4e & 4f). This provides practitioners an alternative to $\mathcal{D}_{\text{overlap}}$ to improve transfer, especially in cases where collecting parallel data is expensive.

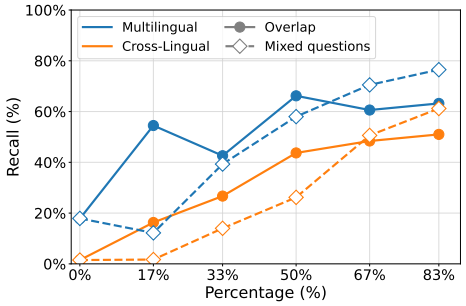


Figure 5: Effect of increasing (a) overlap and (b) number of mixed QA on knowledge recall, averaged across distinct language pairs. XLR scales monotonically with overlap and mixed QA, but with diminishing returns for higher overlap.

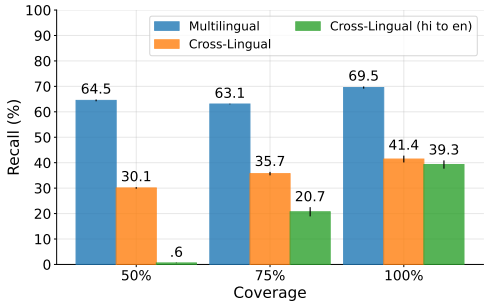


Figure 6: Recall across different overlap methods. Each trades coverage for extent. at the same net overlap. **Higher coverage is more beneficial for XLT, especially for dissimilar languages.** Errors are 2σ deviations from 3 training runs.

The emergence of XLT with dispersion is surprising, as the data is unparallel unlike $\mathcal{D}_{\text{overlap}}$ and does not contain explicit code-mixing like $\mathcal{D}_{\text{mix}}$. We speculate a few reasons for this result. Documents in $\mathcal{D}_{\text{disp}}$ are shorter on average, implying many of them may be packed within a single sequence during training. This increases the language diversity per training batch. Optimizing next-token prediction may encourage learning associations between languages to better predict the frequent language switching, resulting in transfer. We confirm this effect of increase in diversity on transfer using an ablation where data within a batch is restricted to a single language. This drastically reduces transfer from $\mathcal{D}_{\text{disp}}$, but not other recipes like $\mathcal{D}_{\text{overlap}}$. More details are in Appendix §D.6.

5.5 Effects of Recipe Combinations

Combining recipes further improves cross-lingual transfer. Figure 2 also highlights the effectiveness of combining recipes. In general, recipe combinations are more effective for XLT, boosting both MLR and XLR. Different recipes target different aspects of the data and can be combined effectively. For example, mixing operates at the QA level without modifying documents, while the diversity introduced by dispersion is largely preserved under overlap. Thus, different recipes can complement each other and further improve recall. Controlled experiments in Appendix Fig. 7 highlight such complementary effects. For example, combining with $\mathcal{D}_{\text{mix}}$ boosts MLR for `hi`, while also improving the symmetry of transfer (Fig. 7c,7e). A combination of all the three recipes – $\mathcal{D}_{\text{overlap}}, \mathcal{D}_{\text{disp}}, \mathcal{D}_{\text{mix}}$ yields the highest average XLR (Fig. 7f), whereas for multilingual recall and XLR across dissimilar languages, the best strategy combines knowledge overlap with QA mixing (Fig. 7c).

6 Conclusion

Our work provides insights into the efficacies of different data curation recipes for cross-lingual knowledge transfer. We offer clear recommendations for multilingual pre-training that reliably improves model capabilities, rather than the usual unsupervised and noisy process which leads to accidental capabilities. From our findings, we recommend that practitioners curating multilingual corpora (a) focus on higher coverage of documents over extent per document, (b) make use of language-mixed data during pre-training, and (c) curate data to complement information in existing corpora. With such informed data curation, we can hope to build stronger multilingual models.

Limitations

Our work has a few important limitations. We use data from factual sources such as biographies and Wikipedia articles, excluding other forms of content. In real-world corpora, the presence of non-factual data may also affect knowledge transfer. However, precisely quantifying factual versus non-factual data in existing corpora is challenging and outside the scope of this work. We fine-tune models with at most 4B parameters. For larger models, the extent of transfer may differ. Lastly, we explore transfer resulting from three dominant compositions in multilingual corpora, namely overlap, dispersion and mixing. There may exist other compositions which also play a role.

AI USE STATEMENT

We used generative AI tools for data processing and evaluation as a judge. As part of our study, we created synthetic datasets, and used generative tools for their synthesis, validation, and translation. Usage details have been documented in sections §4, and Appendices §A.1, §A.2 and §B.2. We also validate the LLM judges used for our evaluations, described in §B.2. We did not use generative tools for other tasks, including writing, finding relevant literature, ideating, or executing experiments. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the necessary details to replicate our experiments, including hyperparameters, training details and architectural details in §4 and in the Appendices §B.1 and §B.2. All models and datasets used throughout the study have been fully documented in the respective sections. We plan to release the source code and data used to facilitate further research upon publication.

REFERENCES

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models. 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. 5454–5467, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.374. URL https://aclanthology.org/2022.acl-long.374/.

Meta AI. Introducing llama 4: Advancing multimodal intelligence, 2024. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models: Part 3.1, Knowledge Storage and Extraction. *SSRN Electronic Journal*, 2025a. ISSN 1556-5068. doi: 10.2139/ssrn.5250633. URL http://dx.doi.org/10.2139/ssrn.5250633.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws. *SSRN Electronic Journal*, 2025b. ISSN 1556-5068. doi: 10.2139/ssrn.5250617. URL http://dx.doi.org/10.2139/ssrn.5250617.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon Series of Open Language Models, 2023. URL https://arxiv.org/abs/2311.16867.

Verena Blaschke, Masha Fedzechkina, and Maartje ter Hoeve. Analyzing the Effect of Linguistic Similarity on Cross-Lingual Transfer: Tasks and Experimental Setups Matter, 2025. URL https://arxiv.org/abs/2501.14491.

Terra Blevins and Luke Zettlemoyer. Language Contamination Helps Explains the Cross-lingual Capabilities of English Pretrained Models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3563–3574, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.233. URL https://aclanthology.org/2022.emnlp-main.233/.

Carter Blum, Katja Filippova, Ann Yuan, Asma Ghandeharioun, Julian Zimmert, Fred Zhang, Jessica Hoffmann, Tal Linzen, Martin Wattenberg, Lucas Dixon, and Mor Geva. Beyond the Rosetta Stone: Unification Forces in Generalization Dynamics, 2025. URL https://arxiv.org/abs/2508.11017.

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.

Nadezhda Chirkova and Vassilina Nikoulina. Zero-shot cross-lingual transfer in instruction tuning of large language models, 2024. URL https://arxiv.org/abs/2402.14778.

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, 2024.

Leandro De Souza, Thales Almeida, Roberto Lotufo, and Rodrigo Frassetto Nogueira. Measuring Cross-lingual Transfer in Bytes. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7526–7537, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.418. URL https://aclanthology.org/2024.naacl-long.418/.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3610–3623, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.264. URL https://aclanthology.org/2022.naacl-main.264/.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, November 1971. ISSN 0033-2909. doi: 10.1037/h0031619. URL http://dx.doi.org/10.1037/h0031619.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6101–6117, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.339. URL https://aclanthology.org/2024.naacl-long.339/.

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, et al. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

Kaiyu He, Tong Zhou, Yubo Chen, Delai Qiu, Shengping Liu, Kang Liu, and Jun Zhao. Semantic Pivots Enable Cross-Lingual Transfer in Large Language Models, 2025. URL https://arxiv.org/abs/2505.16385.

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, November 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/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, and Fartash Faghri. TiC-LM: A Web-Scale Benchmark for Time-Continual LLM Pretraining. 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. 32231–32273, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1551. URL https://aclanthology.org/2025.acl-long.1551/.

Zeman Li, Yuan Deng, Peilin Zhong, Meisam Razaviyayn, and Vahab Mirrokni. PiKE: Adaptive data mixing for large-scale multi-task learning under low gradient conflicts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=xNJenVNmzL.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schuetze. Tracing Multilingual Factual Knowledge Acquisition in Pretraining. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2121–2146, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. URL https://aclanthology.org/2025.findings-emnlp.113/.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, 2019. URL https://arxiv.org/abs/1711.05101.

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/.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, 2024. URL https://arxiv.org/abs/2406.17557.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language, 2025. URL https://arxiv.org/abs/2506.20920.

Vihari Piratla, Purvam Jain, Darshan Singh, Trevor Cohn, Preethi Jyothi, and Partha Talukdar. Rethinking cross-lingual gaps from a statistical viewpoint, 2026. URL https://arxiv.org/abs/2510.15551.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

Chen Shani, Yuval Reif, Nathan Roll, Dan Jurafsky, and Ekaterina Shutova. The roots of performance disparity in multilingual language models: Intrinsic modeling difficulty or design choices?, 2026. URL https://arxiv.org/abs/2601.07220.

Noam Shazeer. GLU Variants Improve Transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Yingli Shen, Wen Lai, Shuo Wang, Ge Gao, Kangyang Luo, Alexander Fraser, and Maosong Sun. From Unaligned to Aligned: Scaling Multilingual LLMs with Multi-Way Parallel Corpora. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7368–7390, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.374/.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, et al. Gemma 3 Technical Report, 2025. URL https://arxiv.org/abs/2503.19786.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, and et al. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. Investigating and Scaling up Code-Switching for Multilingual Language Model Pre-Training, 2025a. URL https://arxiv.org/abs/2504.01801.

Zining Wang, Yuxuan Zhang, Dongwook Yoon, Nicholas Vincent, Farhan Samir, and Vered Shwartz. WikiGap: Promoting Epistemic Equity by Surfacing Knowledge Gaps Between English Wikipedia and other Language Editions, 2025b. URL https://arxiv.org/abs/2505.24195.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. 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. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do Large Language Models Latently Perform Multi-Hop Reasoning? 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. 10210–10229, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550/.

Çağatay Yıldız, Nishaanth Kanna Ravichandran, Nitin Sharma, Matthias Bethge, and Beyza Ermis. Investigating Continual Pretraining in Large Language Models: Insights and Implications. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=aKjJoEVKgO.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. How Vocabulary Sharing Facilitates Multilingualism in LLaMA? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12111–12130, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.721. URL https://aclanthology.org/2024.findings-acl.721/.

Wenhao Zhuang and Yuan Sun. CUTE: A Multilingual Dataset for Enhancing Cross-Lingual Knowledge Transfer in Low-Resource Languages. 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. 10037–10046, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.670/.

APPENDIX

A Dataset Construction Details **15**

A.1 Synthetic Biographies for Controlled Experiments 15

A.2 Wikipedia Articles for Real-World Experiments 16

A.3 Dataset Statistics 18

B Experimental Details **18**

B.1 Controlled Experiments 18

B.2 Real-World Experiments 19

C Translation Cost of Different Recipes **21**

D Additional Results **21**

D.1 Effects of Combinations of Data Recipes 21

D.2 Real-World Setup Results 21

D.3 Effects on General Multilingual Capabilities 22

D.4 Evaluation on Two-Hop Questions 22

D.5 Exploring the Asymmetry in Transfer across Overlap 24

D.6 Explanation of the Effects of Dispersion 25

E Prompts **26**

A DATASET CONSTRUCTION DETAILS

A.1 SYNTHETIC BIOGRAPHIES FOR CONTROLLED EXPERIMENTS

For studying the role of data compositions, we curate information about individuals in the form of biographies. Our biography data extends the dataset employed by Allen-Zhu & Li (2025a) to study knowledge acquisition in a monolingual setting. For each individual, we generate fictitious data from a pool of $K = 6$ attributes This implies that the number of attributes for each biography is the same, i.e., $K(d) = K$. In our dataset, we generate data for $N = 10^5$ individuals. We treat individual attributes as segments for our data recipes.

Each individual has the following information, sampled in isolation from a uniform distribution:

- **Name**: Names for an individual are created by sampling a first, middle and last name from a pool of 400, 400 and 1000 unique name parts. We employ rejection sampling to ensure unique names.
- **Gender**: We randomly assign each individual a gender from male, female or non-binary. This information is only used while selecting an appropriate sentence template.
- **Birth Date**: Birth date for an individual is sampled from the range 1900-01-01 to 1966-12-31, with $\sim$ 22K unique values. Dates are formatted as DD MMM YYYY, with the month in words.
- **Birthplace**: We sample birthplace for an individual from a pool of 200 unique values. Each location includes an ISO country code.
- **University**: We sample university names from a pool of 300 unique place names merged with a random template.
- **Major**: We sample majors from a pool of 50 unique majors.

- **Employer**: We sample employer names from a pool of 300 unique company names.
- **Employer Place**: We sample employer places randomly for each employer. In our data, we obtain 259 unique places.

We create biographies for individuals using sentence templates corresponding to different attributes (such as {name} was born in {birth_place}) that are populated with the details of the specific individual. Templates are sampled from a pool with 30 templates per attribute, generated in English using an LLM (LLaMA 3.1 Instruct 70B, Grattafiori et al. (2024)).[2] For every attribute in a biography, we randomly sample a template, and populate it with the individual's attribute value. These sentences are then concatenated together in a fixed order to form the biography of the individual. For each individual, we also curate QA pairs corresponding to each attribute, also populated via templates. These QA instances are used to evaluate the knowledge recall (generalization) of the model.

We translate the English biographies and QA pairs to 2 other languages: French, and Hindi. We employ Google Translate to obtain the translations.[3] For attributes that are proper nouns, we perform transliteration, while for other attributes such as majors and the sentence templates, we use translation. We translate the entirety of data into all other languages, obtaining 3×10^5 biographies in total. We construct datasets for different recipes by first obtaining a $\mathcal{D}_{\text{base}}$, then sampling translations as required to obtain $\mathcal{D}_{\text{overlap}}$, $\mathcal{D}_{\text{mix}}$ or $\mathcal{D}_{\text{disp}}$ respectively.

Data Sources and Licenses. For obtaining data pools, we use open-source datasets from GitHub: `sigpwned/popular-names-by-country-dataset`,[4] `fivethirtyeight/data`, [5] and the Faker Python package.[6] The dataset `sigpwned/popular-names-by-country-dataset` is available for public use under the CreativeCommons CC0 license.[7] The dataset `fivethirtyeight/data` is available for use under the CreativeCommons CC-4.0 license.[8] The Faker library is available for use under the MIT license.

A.2 Wikipedia Articles for Real-World Experiments

We use the real-world setup to demonstrate how certain data compositions such as introducing parallel data (overlap) or code-mixing can encourage knowledge transfer for factual recall. Our setup involves training off-the-shelf models on a corpus of Wikipedia articles with new knowledge. Documents on Wikipedia have many facts, which we intend to train the model for. We test the effectiveness of learning by probing for the learnt knowledge using QA pairs associated with each document.

One problem with naively using Wikipedia is that we may risk training the model on knowledge that it may already have been trained for. This is likely as models undergo pre-training on Wikipedia (Grattafiori et al., 2024). Further, for recent open-weight models, the training data is rarely disclosed. In such cases, we may end up with potential evaluation contamination.

To reduce the risks of contamination, we collect Wikipedia articles that have been created beyond a cutoff date. In our dataset, we set this date to January 2024. This lies close to the knowledge cutoff date of Gemma 3 (August 2024), and is far enough to ensure that we can collect a sufficient number of articles for training. However, some of these articles may be a part of the training data for Gemma 3, and many of the newer articles may still talk about old topics.

We apply a few other quality checks to retain useful documents that can be used for assessing factual knowledge acquisition and transfer. We only collect documents which belong to factual categories, such as people, places, events, etc. Also, we only collect documents that have a minimum character and passage count. The latter is useful as our recipes operate on the level of passages in the real-world setup. With these quality checks, we collect $M = 70,000$ Wikipedia articles for our training corpus.

[2] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct/
[3] https://translate.google.com
[4] https://github.com/sigpwned/popular-names-by-country-dataset
[5] https://github.com/fivethirtyeight/data
[6] https://pypi.org/project/Faker/
[7] https://creativecommons.org/publicdomain/zero/1.0/deed.en
[8] https://creativecommons.org/licenses/by/4.0/

Question-Answer Pairs. Similar to the synthetic setup, we test knowledge acquisition by probing the model for the learnt knowledge using question-answer (QA) pairs. A model's performance on QA reliably estimates how well the model can use the knowledge acquired during pre-training. For this purpose of evaluation, we create a pool of questions associated with each article.

For each passage from a training article, we generate questions using a process inspired by Calderon et al. (2026). To obtain questions, we first identify named entities from the target passage using an LLM. These entities form candidate answers for the questions to be generated. By directly selecting entities from the passage, we ensure grounding to the document. From this list of named entities, we prompt the same LLM to select up to four entities. The model is instructed to select entities based on whether they require factual knowledge recall. A passage may contain a lot of entities, and not all candidates are suitable for evaluation. After this selection, we prompt the same LLM to generate a question for every candidate entity selected in the previous step. The model is given the passage, the candidate entity as the expected answer. In addition, we also instruct the model to ensure the question is self-contained and uses a phrasing similar to the passage itself. We use Gemini 3.1 Pro (`gemini-3.1-pro-preview`) using the `google-genai` API as the LLM for these tasks. Prompts for the identification, selection, and generation stages are given in §E.

After the generation process, we obtain a set of candidate QA pairs for every passage per training article. However, not all questions are usable, as questions may be trivially answerable (because of some hints), have multiple potential answers, require knowledge of the context, or may be vaguely stated. To improve the quality of questions, we run a validation and revision pipeline over every question. We run a series of tests, where an LLM judge assesses whether the question is: (a) ambiguous with respect to the source context, (b) imprecisely stated making the expectation unclear, or is (c) trivial to answer by not relying on the specific context of the article. The model is given the source passage and expected answer for reference, and is asked to either try to revise the question if possible or reject it. Questions that do not pass *any* of these tests are discarded. With questions that pass all checks, we probe the base model to assess whether the model can answer these questions. The base model is tested for exact match of the answers at Pass@5 (5 responses are sampled and evaluated at temperature 1). Questions which the base model can answer are also filtered out, as the associated knowledge is not new. The final set of questions includes those which pass all checks and were not answerable by the base model. For such questions, we believe that there are negligible chances of contamination, and any improvements in recall and transfer occur as a result of the continual pre-training process.

Note that the generation and validation process is conducted only in English. We obtain questions in other languages by translation.

Paraphrases. For every article and QA pair, we perform paraphrasing using an LLM. These paraphrases are then added into the training corpus. Prior work suggests that models acquire factual information reliably through repeated exposures to the same information (Allen-Zhu & Li, 2025b). While one way to achieve repetition is to train for many epochs, such an approach is unrealistic at scale. Hence, we resort to semantic duplication instead, in the form of paraphrases. These paraphrases are included only for the documents we choose to include in the training corpus. We collect up to 4 paraphrases for each document *and* QA pair. Paraphrases are also included for the baseline recipe, for a fair comparison. As the cost of generating paraphrases is also non-trivial, we still perform training for multiple epochs, in conjunction with the use of paraphrases. We use Gemma 4 MoE (26B, 4B active parameters) as the model for paraphrasing.[9] The prompts used are listed in §E. For articles, paraphrasing is performed passage by passage.

Translation. Similar to the controlled setup, we perform translation of all articles and QA pairs in advance to facilitate quick experimentation with various recipes. We translate the English articles and QA pairs to 2 other languages: Kannada, and Hindi. We employ a machine translation model, TranslateGemma 27B to obtain these translations.[10] We use this model in favor of Google Translate due to the bulk of translations involved. Translations are performed locally using `vLLM` (Kwon et al.,

[9] https://huggingface.co/google/gemma-4-26B-A4B-it

[10] https://huggingface.co/google/translategemma-27b-it

2023) library for inference.[11] For every article and QA pair, the corresponding paraphrases are also translated and included for training.

Data Sources and Licenses. We curate articles from the English Wikipedia Dump (multiindex stream, dated '20260701'). Data from the dump is available for public use under the CreativeCommons Attribution-ShareAlike CC By-SA 4.0 license.[12]

A.3 Dataset Statistics

Table 1 summarizes the dataset sizes in terms of training tokens for different recipes across the controlled and real-world setups.

Recipe	Controlled Setup	Real-World Setup
Baseline	13M	1.6B
Overlap	26M	2.4B
Dispersion	13M	1.6B
Mixing (QA)	40M	1.7B
Overlap & Mixing	48M	2.5B

Table 1: Token counts across each recipe in the controlled and real-world setups. Counts are computed using the trained tokenizer for the controlled setup, and the tokenizer of the base model (Gemma 3) for the real-world setup.

B Experimental Details

B.1 Controlled Experiments

Model and Training Setup. To evaluate the effects of data compositions on knowledge transfer, we curate pre-training datasets using one of several data recipes. We pre-train language models from scratch over these datasets. These trained models are then evaluated via multilingual and cross-lingual recall (§3). Using the result of a specific model, we quantify the contribution of different recipes towards XLT.

Each model is a dense LLaMA-based decoder-only transformer, with 12 layers and 12 attention heads. We use pre-LayerNorm, RoPE embeddings (Su et al., 2023) and SwiGLU activations (Shazeer, 2020). The hidden dimension of the model is 768, which is projected up to 3072 in the MLP layers. We avoid tying of embedding weights. The total parameter count is $135M$, including embedding layers. Unless mentioned, we employ defaults used in the `LlamaConfig` from the HuggingFace implementation (Wolf et al., 2020). [13] Across all experiments, we use the same model and training hyperparameters.

We also train our own tokenizer to compensate for the domain-specific data. Our tokenizer is a byte-level BPE tokenizer (Sennrich et al., 2016), trained on a variant of our source dataset where multiple paraphrases of each biography are present and names are included in all sentences. We avoid using the TikToken split pattern used for LLaMA 3 when pre-tokenizing the data (Grattafiori et al., 2024), as we found that using this pattern degrades fertility for languages with non-Latin scripts. Starting from bytes, tokens are merged until a word boundary is reached. The resulting tokenizer has a vocabulary size of $14K$ tokens in total.

For training, we randomly shuffle and collate instances from the dataset into sequences of 512 tokens each. We employ mixed pre-training (Allen-Zhu & Li, 2025a), where we pre-train jointly on biographies and QA data. While collating, we ensure that each sequence only includes either biographies or QA data, but not both, to avoid shortcut learning. Data is ordered randomly, so parallel

[11] http://github.com/vllm-project/vllm

[12] https://creativecommons.org/licenses/by-sa/4.0/

[13] We use the following configuration: modeling_llama.py

sentences across languages (corresponding to multiple biographies about an individual in different languages), if present, may be observed across different batches. We do not truncate instances, padding smaller sequences instead. Training is performed until convergence – until the model's recall for training biographies reaches $\simeq 100\%$. We find that all models achieve convergence under $150K$ steps, typically corresponding to ~ 130 epochs. As per the study by Allen-Zhu & Li (2025b), the amount of exposures is sufficient to encode knowledge at the capacity of 1 bit / parameter.

Training Hyperparameters. We use the HuggingFace `accelerate` library (Gugger et al., 2022) to perform pre-training experiments. Prior to training, we identify hyperparameters using a simple hyperparameter sweep over a smaller sample of our data ($N = 10^3$ biographies). Hyperparameters corresponding to the run with the best training loss are then used across the main experiments. We conduct a sweep over the learning rate across $\eta = \{10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$ and batch size $b = \{64, 96, 128\}$. For optimization, we employ AdamW (Loshchilov & Hutter, 2019) with $\epsilon = 10^{-6}$ and a weight decay of 0.1. We train each model for next-token prediction using categorical cross-entropy. Additionally, we regularize the output logits using Z-loss, scaled by a factor of $\lambda = 10^{-4}$, following its effectiveness for training stability (Almazrouei et al., 2023). During training, we employ gradient clipping to ensure the norm of the gradients is below 1.2. From the sweep, we identify $\eta = 10^{-3}$ and $b = 96$ as the choice of hyperparameters. We tune the learning rate using a cosine schedule with warmup, where η is first increased to 10^{-3} across the first $1K$ steps, and then is decayed via a cosine schedule to 10^{-4}. Training is performed in mixed `bfloat16` precision. Details are summarized in Table 2.

Architecture		**Training**	
Detail	**Value**	**Detail**	**Value**
Layers	12	Optimizer	AdamW
Attention heads	12	Learning rate	10^{-3} (cosine)
Hidden dimension	768	Batch size	96
MLP dimension	3072	Weight decay	0.1
Parameters	135M	Grad clip norm	1.2
Positional encoding	RoPE	Z-loss scale	10^{-4}
Activation	SwiGLU	Sequence length	512 tokens
Embedding tying	No	Max steps	150K
Tokenizer	Byte BPE (∼14K)	Precision	`bf16`

Table 2: Architecture and training configuration summary for controlled experiments.

For each configuration, we perform 3 runs over the same dataset, with different random seeds that control weight initialization, and dataset shuffling, and report the average results of the best checkpoint for each run.

Compute. All experiments are conducted over a single NVIDIA H200 GPU. Each training run takes ~ 8 hours to complete.

Evaluation. Post training, we perform evaluation over the best checkpoint, determined using a held-out sample of the evaluation set $\mathcal{Q}^{(\text{te})}$ reserved for validation. During evaluation, we sample responses using beam search for decoding with a beam size of 5. We sample at most 20 tokens. A response is correct if it *exactly matches* the expected answer. For a fair comparison of recipes, we report evaluation scores over the intersection of test sets $\mathcal{Q}^{(\text{te})}$ of each recipe being compared. This intersection set is further stratified by language and attribute to reduce the effects of imbalance of either of these factors in the evaluation set. Inference is performed using the `transformers` (Wolf et al., 2020) library.

B.2 Real-World Experiments

Model and Training Setup. To validate the effects of data compositions on knowledge transfer, we curate datasets using one of several data recipes. We continually pre-train an off-the-shelf language

model over these datasets. These trained models are then evaluated via multilingual and cross-lingual recall (§3). Using the result of a specific model, we validate the contribution of different recipes towards XLT. We experiment with the Gemma 3 model (Team et al., 2025), with 4B parameters. [14]

For training, we randomly shuffle and collate instances from the dataset into sequences of 16K tokens each. We employ mixed pre-training as in the controlled setup, where we pre-train jointly on biographies and QA data. While collating, we ensure that each sequence only includes either biographies or QA data, but not both, to avoid shortcut learning. We do not truncate instances, padding smaller sequences instead. All articles in the corpus fit within the 16K context window that we use for training. Training is performed for 3 epochs. Across all experiments, we use the same model and training hyperparameters.

Training Hyperparameters. We use the HuggingFace `accelerate` (Gugger et al., 2022) and `unsloth` libraries to perform continual pre-training experiments.[15] For optimization, we employ AdamW (Loshchilov & Hutter, 2019) with a weight decay of 0.01. We train each model for next-token prediction using categorical cross-entropy. We use a smaller learning rate of $\eta = 10^{-4}$ and batch size of 64 as the choice of hyperparameters. For the embedding matrices, a smaller learning rate of 5×10^{-5} is used, to reduce forgetting. We tune the learning rate using a cosine schedule with warmup, where η is first increased to 10^{-4} across the first 10 steps, and then is decayed via a cosine schedule to 5×10^{-5}. Training is performed in mixed `bfloat16` precision.

Compute. All experiments are conducted using $2\times$ NVIDIA H200 GPUs. Each training run takes at most $\sim$ 29 hours to complete, depending on the model and data recipe. Due to limitations in obtaining compute for prolonged periods, we only conduct single training runs per data recipe. When reporting results, we show variance through bootstrapped samples.

Evaluation. Post training, we perform evaluation over the best checkpoint, determined using a held-out sample of the evaluation set $\mathcal{Q}^{(\text{te})}$ reserved for validation. During evaluation, we sample responses using greedy decoding, sampling at most 96 tokens. We also pass a few examples of QA pairs sampled from the training set $\mathcal{Q}^{(\text{tr})}$ for in-context learning in a few-shot prompt, as we work with base models. We use the same few-shot examples across all evaluations. Examples are sampled randomly without considering the output language or topic. For a fair comparison of recipes, we report evaluation scores over the intersection of test sets $\mathcal{Q}^{(\text{te})}$ of each recipe being compared. This intersection set is further stratified by language to reduce the effects of imbalance in the evaluation set. Inference is performed using the `unsloth` library.

Unlike the controlled setting, models may produce semantically equivalent responses which may not exactly match the expected answer. As such, we employ an LLM as a judge to rate responses, using the approach in Wei et al. (2024). The judge accepts a model response and an expected answer, and assigns a grade between 'A', 'B', and 'C' to denote if the response is correct, incorrect, or unattempted, respectively. We use Gemma 4 MoE (26B, 4B active parameters) as the judge for our experiments. The prompt for the judge is described in §E. Inference for the judge is performed using `vLLM` (Kwon et al., 2023). In our computation of recall, we treat a grade of correct ('A') as a correct response, and other grades as incorrect.

Choice of Evaluation Judge. We also experiment with a few other judges, namely Llama 4 Scout (AI, 2024) [16] and Qwen 3.6 27B (Qwen Team, 2026), [17] to ensure that our evaluation is free of any bias that may arise from evaluating models using another model from the same family. We measure inter-annotator agreement between the judges, using Fleiss Kappa, κ (Fleiss, 1971). Cross-lingual recall scores for the models we train as evaluated by the different judges are given in Table 3.

We find high agreement amongst the different judges we explore, implying that the trends we observe are independent of our design choices for evaluation. As such use Gemma 4 MoE for our final experiments.

[14] https://huggingface.co/google/gemma-3-4b-pt

[15] https://github.com/unslothai/unsloth

[16] https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct

[17] https://huggingface.co/Qwen/Qwen3.6-27B

Recipe	Evaluation Judge			Agreement
	Gemma 4 MoE (*)	Llama 4 Scout	Qwen 3.6	Fleiss κ
Baseline	6.3	5.2	7.7	0.9
Overlap	15.3	14.2	16.8	0.9
Dispersion	11.4	10.5	13.0	0.9
Mixing (QA)	10.6	9.7	11.8	0.9
Overlap & Mixing	18.4	17.1	19.9	0.9

Table 3: Average cross-lingual recall (XLR, in %) for real-world experiments as evaluated by different judges. Asterisk (*) denotes the default judge originally used for evaluation. All judges show high levels of agreement, validating our evaluation.

C Translation Cost of Different Recipes

In this section, we discuss the translation cost of various recipes. We assume that the cost of translating a segment is constant and same as that of translating a QA pair. We calculate translation costs in units of this fixed constant cost.

Baseline ($\mathcal{D}_{\text{base}}$) This recipe has no translation cost, since it is an unparallel corpus of documents.

Knowledge Overlap ($\mathcal{D}_{\text{overlap}}$) For coverage f and extent p, the cost of curating $\mathcal{D}_{\text{overlap}}$ is:

$$\sum_{d \in \mathcal{D}_{\text{base}}} \underbrace{\lfloor f \cdot N \rfloor}_{\text{\# documents}} * \underbrace{\lfloor p \cdot K(d) \rfloor}_{\text{\# segments per document}} * \underbrace{(\mid \mathcal{L} \mid -1)}_{\text{\# languages}}$$

The cost of translating the associated QA data is the same.

Knowledge Dispersion ($\mathcal{D}_{\text{disp}}$) The cost of curating $\mathcal{D}_{\text{disp}}$ is the same as $\mathcal{D}_{\text{overlap}}$ as the former translates the same amount of segments as the latter, with the only difference being that in curating $\mathcal{D}_{\text{disp}}$ the segments are removed from the original document. In practice, however, the cost of curating such data composition may be zero, since such arrangements are known to exist naturally (Wang et al., 2025b).

Language Mixing ($\mathcal{D}_{\text{mix}}$) When applying mixing to documents, the translation cost can be assumed to be negligible since only the entity name tokens are translated. Similarly, when applying the recipe to QA data, the practitioner must only translate a handful of answer tokens. Therefore, the overall translation cost involved here is negligible.

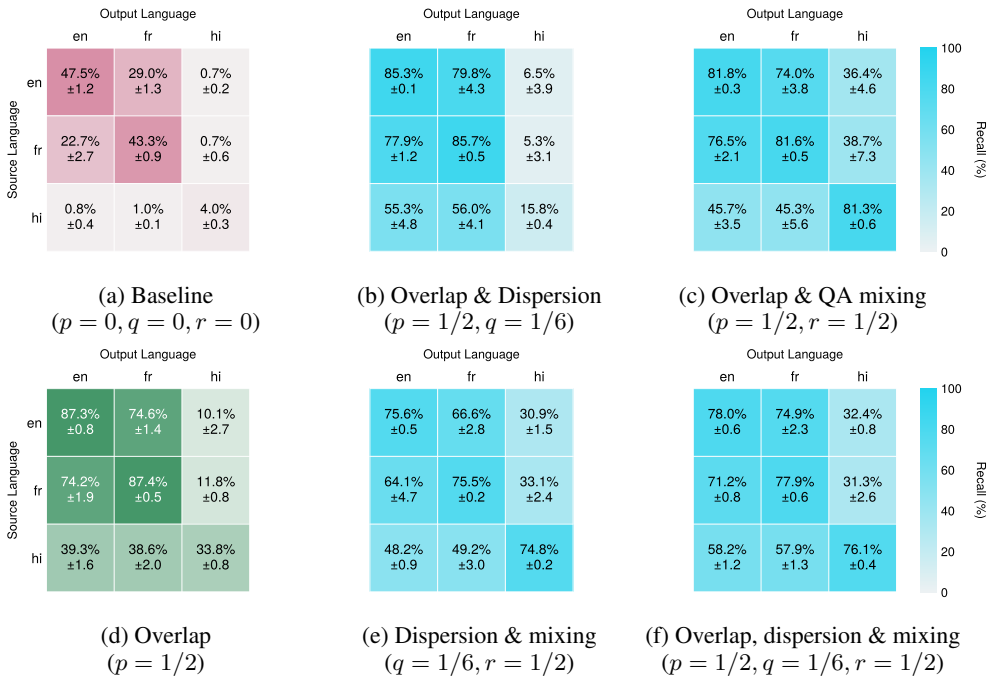
D Additional Results

D.1 Effects of Combinations of Data Recipes

Figure 7 compares the effects of combinations of data recipes on cross-lingual transfer (XLT). Different recipes complement each other positively, significantly improving XLT on average.

D.2 Real-World Setup Results

Fig. 8 provides a breakdown of multilingual and cross-lingual recall across languages for experiments where we continually pre-train Gemma 3 4B (Team et al., 2025) as per our data recipes. We find that data recipes effective in the controlled setting reliably improve knowledge transfer in this real-world setting. Overlap yields the highest improvements, up to $\sim 8\%$ above the baseline (Fig. 8c). Recipe combinations are even more effective, complementing the effects of individual recipes and further improving XLT (Fig. 8f).



(a) Baseline ($p = 0, q = 0, r = 0$)

(b) Overlap & Dispersion ($p = 1/2, q = 1/6$)

(c) Overlap & QA mixing ($p = 1/2, r = 1/2$)

(d) Overlap ($p = 1/2$)

(e) Dispersion & mixing ($q = 1/6, r = 1/2$)

(f) Overlap, dispersion & mixing ($p = 1/2, q = 1/6, r = 1/2$)

Figure 7: Effect of combinations of data recipes on cross-lingual recall. Unless stated, mixing denotes mixing across QA pairs. Different methods complement each other, significantly boosting XLT. Errors denote 2σ deviations from 3 training runs.

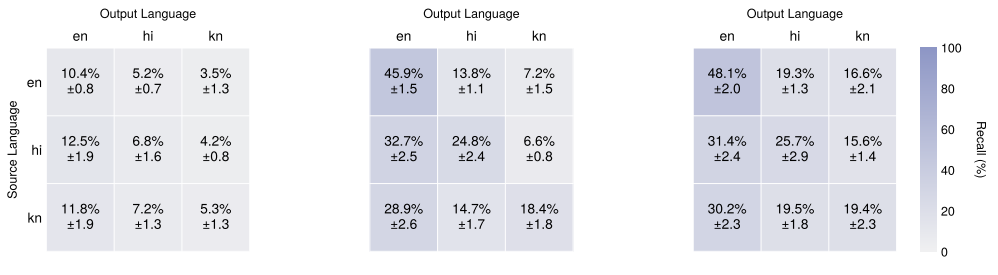
D.3 Effects on General Multilingual Capabilities

Our experiments in the real-world setting continually pre-train off-the-shelf language models over a corpus of Wikipedia articles. Data recipes that we study alter the composition of this corpus, particularly by adding more data. While these recipes benefit cross-lingual transfer, they may adversely affect general multilingual capabilities. To test whether this is truly the case, we compare bits-per-byte (BPB) scores of models that result from different data recipes. BPB is defined as the average loss of the model (in bits) per data byte, over a held-out set of documents. We compare BPB instead of perplexity, as the former is independent of the fertility of the tokenizer for a given language, which may lead to over-estimation for low-resource languages with poor fertility (Shani et al., 2026). For each model, we compute BPB over the languages we train on, namely English, Hindi, and Kannada. We also compute BPB for French, a language which we do not train on, to assess effects of training on other languages. We use a random sample of 300 documents from FineWeb (Penedo et al., 2024) and FineWeb2 (Penedo et al., 2025) corpora to measure BPB in English and other languages, respectively. When computing BPB, we estimate the total loss over a sequence using a sliding window approach, where we use a maximum context length of 16K tokens, and a sliding window of 256.

Table 4 summarizes the BPB scores for trained models resulting from different recipes across different languages. We find that in general, while BPB increases due to training, the relative changes across different recipes are small. This suggests that data recipes that we study do not affect general multilingual capabilities, beyond the effects of continual pre-training over a domain-specific corpus (such as Wikipedia articles in our case).

D.4 Evaluation on Two-Hop Questions

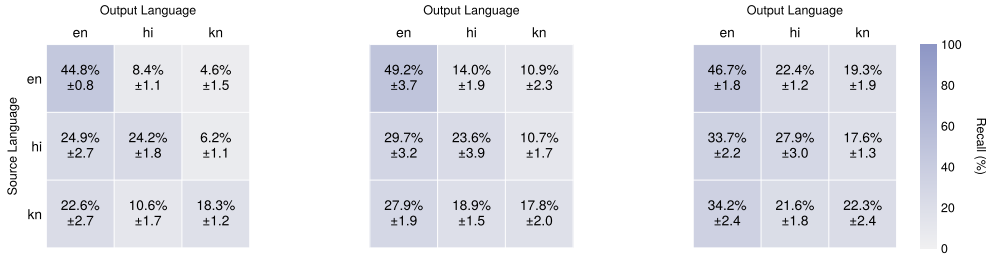
In our main experiments, we evaluate Gemma 3 models continually pre-trained on Wikipedia articles for the task for factual recall in a language other than the source. Using this evaluation setup, we find that our data recipes incentivize transfer beyond that of the source. However, factual recall is only one aspect of knowledge transfer, and it is unclear whether the knowledge transferred can be utilized



(a) Base Model (before training)

(b) Mixing across QA ($r = 1/2$)

(c) Knowledge overlap ($p = 1/2$)



(d) Baseline ($f = 0$)

(e) Knowledge dispersion ($q = 0.3$)

(f) Overlap, & QA mixing ($p = 1/2, r = 1/2$)

Figure 8: Effect of different data recipes on cross-lingual recall when continually pre-training Gemma 3 (4B). Recipes effective in the control setting also improve XLT in the real-world setting. Errors denote 2σ deviations from 200 bootstrap samples.

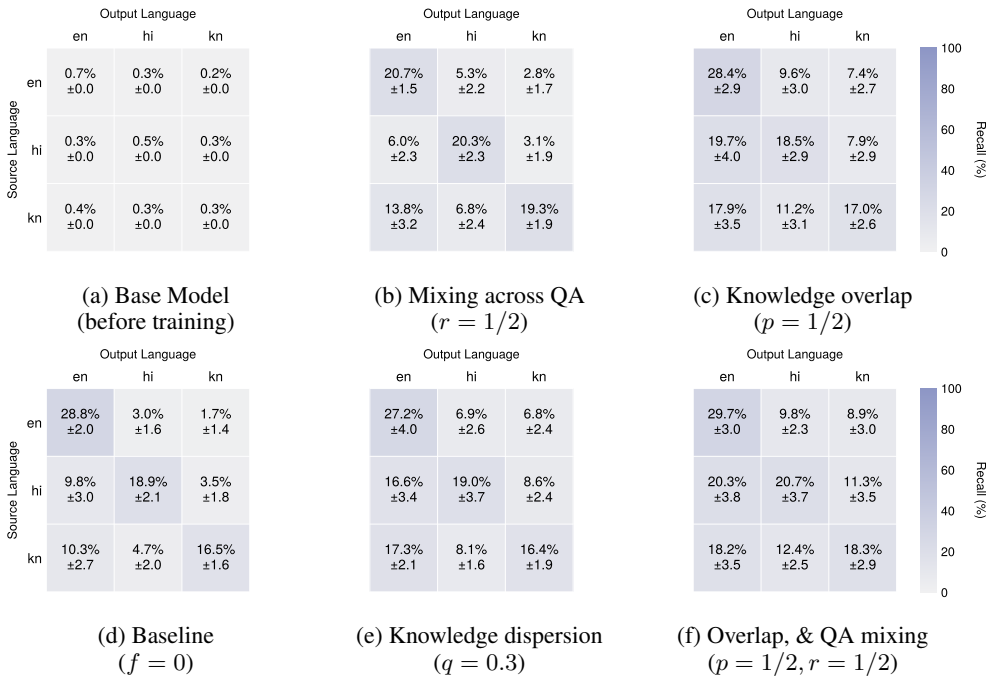
Recipe	en	kn	hi	fr
Base Model	0.82	0.42	0.36	0.79
Baseline	1.09	0.65	0.57	1.08
Overlap	1.10	0.67	0.61	1.13
Dispersion	1.10	0.71	0.59	1.11
Mixing (QA)	1.11	0.66	0.59	1.12
Overlap & Mixing	1.16	0.68	0.62	1.19

Table 4: BPB scores across recipes and languages for Gemma 3 4B. Across recipes, the effects of training on BPB scores are similar.

effectively across other contexts. To understand whether the knowledge transferred is effective beyond simple recall and whether models generalize to other tasks, we evaluate the Gemma models we train on a harder task of multi-hop questions. Such questions test a mix of both factual recall and reasoning capabilities, as answering such questions correctly requires both recalling knowledge of the underlying hops and the ability to reason and bridge the hops to deduce the answer.

We evaluate the ability of models to reason over the learnt knowledge using two-hop questions. We construct these questions by merging pairs of single-hop questions from our unseen evaluation set. These questions are also generated using Gemini 3.1 Pro, using a similar process as in §A.2. To merge questions, we first identify compatible candidate pairs across a document, then generate merged questions for each pair. Generated questions undergo a similar process of validation to ensure unambiguity, preciseness and grounding. In addition, we also validate that the questions include a genuine two-hop. We also filter out any questions answerable by the base model. The remaining questions are then used for evaluation. Prompts to generate these questions are also given in §E.

For evaluation, we measure the model's cross-lingual recall over the two-hop questions, using the same process for evaluation as for the single-hop questions (refer §B.2 for details). We also quantify the fraction of multi-hop questions for which both the source hops are correctly answered (and are



(a) Base Model (before training)

(b) Mixing across QA ($r = 1/2$)

(c) Knowledge overlap ($p = 1/2$)

(d) Baseline ($f = 0$)

(e) Knowledge dispersion ($q = 0.3$)

(f) Overlap, & QA mixing ($p = 1/2, r = 1/2$)

Figure 9: Cross-lingual recall on two-hop questions post continually pre-training Gemma 3 (4B) with different data recipes. Recipes that improve single-hop recall also improve multi-hop performance. Errors denote 2σ deviations from 200 bootstrap samples.

thus known) in the same language as the question. The latter represents an ideal upper bound on the expected XLR across two-hop questions. Results are shared in Figure 3.

Fig. 9 provides a breakdown of multilingual and cross-lingual recall across languages for Gemma 3 4B (Team et al., 2025) over two-hop questions. Recipes that improve single-hop recall show similar improvements across multi-hop performance. Recipe combinations continue to be the most effective, yielding the highest XLR.

We learn that recipes that improve single-hop recall also perform better on multi-hop questions. Better recall increases the fraction of two-hop questions for which both hops are known, potentially leading to improved multi-hop performance. While this does not show that transfer improves reasoning capabilities, this does suggest that knowledge transferred across languages can be effectively reasoned about in downstream tasks. Thus, knowledge transferred across different recipes is effective beyond simple recall.

D.5 Exploring the Asymmetry in Transfer across Overlap

Across the main results in §5.2, we observe that transfer from knowledge overlap is asymmetric across dissimilar language pairs. For example, transfer from `en` to `hi` is only 10.1%, while transfer from `hi` to `en` is 39.3% (Fig. 5.2). This is surprising, given that in our controlled experiments, we explicitly control for language data imbalance and use equal amounts of data across all languages (§4). We also observe that across pairs of languages with asymmetric transfer, multilingual recall (MLR) in the source language is also poor. For example, across knowledge overlap we observe high recall in `en` (up to 87.3%), but significantly lower recall in `hi` (only 33.8%). Across other cases where MLR in `hi` is comparable to `en`, such as in the case of combining overlap with QA mixing where MLR in `en` is 81.8% and 81.3% in `hi`, we also observe that the transfer between `hi` and other languages is more balanced (`en` $\rightarrow$ `hi` at 36.4% versus `hi` $\rightarrow$ `en` at 45.7%, Fig. 7c). This suggests that poor MLR in a language and asymmetry in transfer to and from that language are related phenomena.

We list down a few simple hypotheses to explain the origins of this asymmetry and the poor multilingual recall. However, as we explore, these potential hypotheses are unable to reliably explain the origin of this asymmetry.

Data Imbalance due to Tokenization. While our setup ensures that the number of biographies and QA instances we train on are exactly the same across all languages, the number of tokens across languages may not be the same. This may arise from a relatively poor fertility of the tokenizer in hi, as en and fr share many tokens. This may bias the model towards certain languages. However, we find that the tokenizer we train exhibits comparable fertility across the three languages. Notably the fertility in hi (1.27) is comparable to that of en (1.16) and fr (1.24). Thus, we obtain a similar number of tokens across all languages.

Data Imbalance due to Shared Script. Despite an equal number of tokens across every language, we may still end up with a data skewness, due to our choice of languages. We experiment with three languages, two of which share a script. As such, the number of tokens in the Latin script are double that of the script of hi. However, we find that this skewness alone does not explain the asymmetry we observe. We run an ablation where we train models with only two languages, fr and hi, eliminating the data skew. Even in this setting, we continue to observe an asymmetry in XLR, where transfer from fr $\rightarrow$ hi is only 3.2%, while transfer from hi $\rightarrow$ fr is 11.9%. We also observe widely different values for MLR, at 49.2% for fr and 3.8% for hi.

Higher Response Variance across Languages. Recent work finds that cross-lingual gaps may emerge from an increase in response variance across target languages (Piratla et al., 2026). If the models' responses have a high variance in certain languages like hi, this may decrease XLR from other languages to hi, leading to the observed asymmetry. Higher variance in hi may arise due to the imbalance created by the shared scripts of en and fr, where most entities are written and tokenized in the same way. To verify if this is the case, we compare XLR@5 scores across language pairs. We sample 5 responses from the model at temperature 1 and assign a correct score if *any* of the responses matches the ground truth. Comparing transfer between en and hi, we find no significant improvements in XLR@5: en $\rightarrow$ hi improves from 10.1% to 12.0%, while hi $\rightarrow$ en increases from 39.3% to 40.7%. Notably, asymmetry continues to exist between the two languages, suggesting that response variance does not characterize the observed asymmetry.

Overall, we find that simple explanations are insufficient to explain the asymmetry in transfer across dissimilar language pairs. We leave further investigation of this asymmetry to future work.

D.6 Explanation of the Effects of Dispersion

We hypothesize that dispersion improves knowledge transfer due to the increased language diversity within a batch. We refer to language diversity as the number of times a change in language is encountered across different sequences within a batch. As documents in dispersion are smaller than usual, many can be packed into a single sequence as per our training process. This leads to an increase in language diversity across a batch on average. Table 5 summarizes batch-wide average statistics for a few of the data recipes we consider. From the table, we confirm that dispersion increases language diversity.

We further hypothesize that an increase in language diversity improves transfer, due to the underlying training objective. Optimizing next-token prediction over sequences with higher diversity may encourage learning associations between languages, to better predict the frequent language switching. [18] This may have similar effects as mixing, resulting in transfer.

To confirm this effect, we run an ablation in our controlled setup, where we intervene to eliminate language-switching within a batch. In this ablation, the data is shuffled in a manner that ensures that every training batch has instances only in a single language. In the presence of this intervention, we re-train models in our controlled setup for two data recipes, overlap and dispersion, and compare the cross-lingual recall before and after the intervention. Table 6 summarizes the results. We find that while for overlap the change is marginal (41.4% to 39.8%), the change in XLR across dispersion is significantly larger (23.7% to 9.9%), approaching that of the baseline. This suggests that language diversity facilitates knowledge transfer and contributes dominantly for knowledge dispersion, but not

[18] Higher diversity within a batch is also known to be important in general for multitask learning (Li et al., 2026). Viewing languages as 'tasks', the increased diversity may also explain the transfer result.

Recipe	Biographies	Language Switches
Baseline	298	165
Overlap	358	211
Dispersion	446	280

Table 5: Batch-wide statistics for data recipes in the controlled setup. Numbers denote the average values across batches encountered during training. Dispersion significantly increases both the number of observed instances and the language diversity per batch on average.

	Cross-Lingual Recall (%)	
Recipe	Before	After
Baseline	9.2	8.9
Overlap	41.4	39.8
Dispersion	23.7	9.9

Table 6: Average cross-lingual recall for data recipes in the controlled setup before and after the intervention to eliminate language-switching within batches. Transfer across dispersion is affected the most, while other recipes exhibit minimal degradation in transfer.

for recipes such as knowledge overlap, where other factors play a larger role. Together, these findings suggest that dispersion benefits transfer from an increase in language diversity.

E PROMPTS

Prompt 1: Prompt for Annotating Entities

```
You will be provided with two parts of a paragraph.
Your task is to perform Named Entity Recognition on the *second*
    part of the paragraph.
Please follow these instructions carefully.

### **Instructions**

1. Read the two parts provided at the end of this prompt.
2. From the second part, identify and select up to a maximum of **
    eight** entities.
3. Do not select lengthy entities (more than five words) or
    entities that are not named entities.
4. Select specific core entities and avoid overly descriptive
    phrases that combine multiple separate concepts (e.g., select "
    documentary photographer" and not "feminist social documentary
    photographer").
5. For each entity you select, you must assign an entity type from
    the following specific list.
6. You should return a new version of the second part of the
    paragraph. Copy-paste the original second part and replace the
    entities you chose with the format `{{{{entity_name; entity_type
    }}}}`
7. Do not replace entities from section headers. These are present
    between '=' symbols.

____

### **Entity Types:**

* **PERSON:** An individual human, whether real or fictional.
* **ORGANIZATION:** A structured company, institution, or
    collective with a common purpose.
* **LOCATION:** A specific geographical place, region, or feature
    on Earth.
* **EVENT:** A notable occurrence or incident that happens at a
    particular time and place.
```

```
* **WORK_OF_ART:** A specific creative work, such as a book, movie,
    song, or painting.
* **PRODUCT:** A commercially produced good, food, service, or
    piece of technology.
* **ORGANISM:** A non-human biological life form, including animals,
    plants, and microbes.
* **SUBSTANCE:** A material with a distinct physical or chemical
    composition.
* **DATE:** A specific point in time, including full dates, years,
    or historical periods.
* **NUMERIC_VALUE:** A number representing a specific quantity,
    measurement, or other quantifiable value.
* **CONDITION:** A specific state of being, typically a medical,
    physical, or operational status.
* **TITLE:** An official position, role, award, or honorific held
    by a person or group.
* **LANGUAGE:** Natural languages, dialects, writing systems,
    including phrases.
* **GROUP:** A collective of people defined by a shared nationality,
    ethnicity, or cultural affiliation.
* **LAW:** A specific, formally enacted rule, regulation, or
    statute created by a governing body.
* **IDEA:** An abstract concept, belief system, named phenomenon,
    or cultural practice.
* **CLASSIFICATION:** A specific type, genre, format, class or list
    used to categorize other
entities.
* **IDENTIFIER:** A unique name, code, or alias used as a formal
    reference or label for another entity.
* **OTHER:** A catch- all category for named entities that do not
    fit into any of the other defined types.

----

**Example:**

*Part 1:*
...
*Part 2*
'Liam Tarquin Broady (BROH-dee; born 4 January 1994) is a British
    professional tennis player who competes mainly on the ATP
    Challenger Tour. He reached a career high ranking of world No.
    93 on 25 September 2023.'

*Correct Output:*
{{{{Liam Tarquin Broady; PERSON}}}} (BROH-dee; born {{{{4 January
    1994; DATE}}}}) is a {{{{British; GROUP}}}} professional {{{{
    tennis player ; TITLE}}}} who competes mainly on the {{{{ATP
    Challenger Tour; EVENT}}}}. He reached a career high ranking of
    world No. {{{{93; NUMERIC_VALUE}}}} on 25 September 2023.

*Explanation:*
1. **Liam Tarquin Broady**: This is the full name of an individual
    human, therefore , it as a **PERSON**.
2. **4 January 1994**: This is a specific point in time, a full **
    DATE**.
3. **British**: This word describes a nationality or cultural
    affiliation. According to the definitions, this is a **GROUP**.
4. **tennis player**: This is a title describing a person 's
    profession or role. The correct type is **TITLE**.
5. **ATP Challenger Tour**: This is a named series of tennis
    tournaments, which is a type of recurring **EVENT**.
```

```
6. **93**: This is a number representing a specific quantity (a
    world ranking), therefore, it is a **NUMERIC_VALUE**.

____

### **Required Output Format:**

Copy-paste the second part of the paragraph. In its new version,
    replace the entities you chose with the format `{{{{entity_name;
    entity_type}}}}`, for example: {{{{tennis player; TITLE}}}}. The
     `entity_name` must be exactly as it appears in the original
    text.

RETURN THE OUTPUT ONLY IN THE FORMAT ABOVE. DO NOT INCLUDE ANYTHING
     ELSE.

**Instruction End**

**Please process the following paragraph and identify up to eight
    entities:**

*Part 1:*
{0}
*Part 2:*
{1}
```

Prompt 2: Prompt for Selecting Identified Entities

```
You will be provided with a paragraph annotated with entities in
    the format of ` {{{{entity_name}}}}`.
Your task is to select up to {n_entities} **OBJECT** entities for
    the pre-training completion task (causal language modeling).
Each OBJECT entity should complete a natural "fill-in-the-blank"
    based on the given text: the *left-context*.
Specifically, the left-context should clearly indicate **what kind
    of entity ** (e.g., a person, country, or biological species) is
     expected at that position.
However, correctly identifying the **specific entity value** (e.g.,
     "South Africa", "banana plants") should require recalling a
    world fact.
These selected entities serve as factual anchors for generating our
     suite of completion-based and QA-based robustness tests.

### **Instructions for selecting suitable OBJECT entities**

* Select OBJECT entities that involve recalling factual knowledge.
* The left-context (the text that precedes the entity) must suggest
     **what kind** of entity is expected (e.g., a date, a country, a
     person's name, a title, a profession, a species, an event, a
    number, etc).
* The left-context should contain enough information to strongly
    suggest a single, unique, unambiguous completion.
* OBJECT entities whose specific values are too easy to guess are
    considered poor choices.
  * *Example:* `Airlink is based in Johannesburg, {{{{South Africa
    }}}}` -> {{{{South Africa}}}} is a poor choice because it can be
     easily guessed from the mention of Johannesburg, without
    requiring any knowledge about Airlink.
* The OBJECT should have a strong relationship with an entity
    mentioned earlier in the text.
```

```
  * *Example:* `{{{{Panama disease}}}}... infects {{{{banana plants
    }}}}` -> Knowing what Panama disease is makes "banana plants" a
    highly predictable object of the verb "infects".
* Do not select OBJECT entities from the start of a sentence.
    Ideally, the OBJECT entity should appear at the end of a
    sentence or complete clause.
* Ensure the left-context provides sufficient disambiguation to
    yield a unique and specific completion. If the blank could
    plausibly be filled by multiple entity types, it is not a
    suitable choice.
  * *Example:* The left-context "Michael Jordan was born on..." is
    ambiguous, as multiple people share this name (e.g., the
    basketball player and the ML researcher). In contrast, "Michael
    Jeffrey Jordan was born on..." and "Michael Jordan, the famous
    basketball player, was born on..." clearly indicate a single
    referent, enabling a unique factual prediction.
  * *Example:* `Liam Tarquin Broady is a...` -> This is a poor
    choice. The context is too broad and could be completed with
    multiple entities, such as a profession (tennis player),
    nationality (British), or a descriptor (left-handed).
* The immediate context preceding the entity should constrain the
    expected entity type, strongly signaling what kind of entity is
    likely to follow. The OBJECT entity can follow a preposition
    that narrows down the possibilities.
 * *Example:* `... born on {{{{4 January 1994}}}}` -> The phrase"
    born on" strongly suggests a date.
 * *Example:* `... caused by the fungus {{{{Fusarium oxysporum f.
    sp. cubense}}}}` -> The phrase "the fungus" clearly indicates
    that a specific fungal name should follow.
 * *Example:* `The Virginia Plan (also known as the {{{{Randolph
    Plan}}}})...`-> The phrase" also known as "implies an
    alternative name is expected.
* Do not consider any information that appears after the OBJECT
    entity when making your selection.
* If the uniqueness or the disambiguation of the entity type relies
     on information that appears after the entity, it is a poor
    OBJECT choice.
 * * Example:* `It is their first album to feature {{{{Matt Wachter
    }}}} on bass.` -> {{{{Matt Wachter}}}} is a poor choice. The
    phrase "It is their first album to feature..." does not
    constrain the prediction. The disambiguating detail" on bass"
    comes* after* the entity, which violates the instruction that
    the** OBJECT** must be inferable solely from the left-context.
* You must select up to {n_entities} suitable OBJECT entities and
    return them in a list, sorted from the most to the least
    suitable.
* Ensure that the list uses the exact same form of the entity as it
     appears in the paragraph (enclosed in `{{` and `}}`).

---

### **Examples**

### **Text 1**
{{{{Panama disease}}}} (or Fusarium wilt) is a plant {{{{disease
    }}}} that infects {{{{banana plants}}}} ({{{{Musa spp.}}}}). It
    is a wilting disease caused by the fungus {{{{Fusarium oxysporum
     f. sp. cubense}}}} (Foc). During the{{{{1950s}}}}, an outbreak
    of Panama disease almost wiped out commercial {{{{Gros Michel
    }}}} banana production.
```

```
* {{{{Panama disease}}}} - It appears in the first clause of the
    first sentence and is therefore ineligible. This choice is **
    poor**.
* {{{{disease}}}} - This entity does not require recalling a world
    fact; it can be predicted from the context clues "Fusarium wilt"
     and "infects." This choice is **poor**.
* {{{{Musa spp.}}}} - While recalling the scientific name for
    banana plants is a good test of knowledge, the use of
    parentheses makes the answer less certain. The context could
    also imply other related species. This choice is **good**.
* {{{{banana plants}}}} - The context "a plant disease that infects
    ..." combined with the subject "Panama disease" creates a very
    strong expectation for this specific plant, a well- known fact.
    This choice is **excellent**.
* {{{{Fusarium oxysporum f. sp. cubense}}}} - The preceding phrase
    "... caused by the fungus..." requires a specific, factual
    answer. Knowing the fungus that causes Panama disease is a clear
     test of world knowledge. This choice is **excellent**.
* {{{{1950s}}}} - The phrase "During the..." is too vague. It could
     be completed with many different time periods, making the
    answer unpredictable. This choice is **poor**.
* {{{{Gros Michel}}}} - The context "... wiped out commercial..."
    strongly points to a specific type of banana. However, it doesn'
    t uniquely identify the variety without further knowledge, so it
    's not perfectly constrained. This choice is **good**.

### **Text 2**
Liam Tarquin Broady (BROH-dee; born {{{{4 January 1994}}}}) is a
    {{{{British}}}} professional {{{{tennis player}}}} who competes
    mainly on the {{{{ATP Challenger Tour}}}}. He reached a career
    high ranking of world No. {{{{93}}}} on 25 September 2023.

* {{{{4 January 1994}}}} - The entity appears in the first clause
    of the first sentence, making it ineligible per the instructions
    . This choice is **poor**.
* {{{{British}}}} - The left-context "... is a..." is too generic
    and does not sufficiently narrow down the possibilities. This
    choice is **poor**.
* {{{{tennis player}}}} - The context "... is a British
    professional..." strongly implies a profession. Knowing who Liam
     Broady is leads to a single, specific prediction. This choice
    is **excellent**.
* {{{{ATP Challenger Tour}}}} - The context "who competes mainly on
     the..." points to a tournament or tour. However, the word "
    mainly" introduces uncertainty, meaning the answer is not
    guaranteed to be unique. This choice is **good**.
* {{{{93}}}} - The preceding phrase "... career high ranking of
    world No..." creates a very specific slot for a number.
    Recalling a player's exact ranking is a great test of factual
    knowledge. This choice is **excellent**.

### **Text 3**
Happiness for Beginners is a {{{{2023}}}} American {{{{romantic
    comedy}}}} film starring {{{{Ellie Kemper}}}} and {{{{Luke
    Grimes}}}}, an adaptation of the {{{{Katherine Center}}}} 2015
    novel of the same name. The novel was adapted for the screen and
     directed by {{{{Vicky Wight}}}}.

* {{{{2023}}}} - This entity appears in the first clause, making it
     ineligible. Additionally, the context "... is a..." is too
    general to uniquely predict a year. This choice is **poor**.
* {{{{Ellie Kemper}}}} - The phrase "film starring" clearly points
    to an actor. However, since a film can have multiple stars, the
```

```
    specific actor named first is not uniquely predictable. This
    choice is **good**.
* {{{{Luke Grimes}}}} - The context "... starring Ellie Kemper and
    ..." creates a strong expectation for a co-star, a common
    structure in this genre. This requires specific knowledge about
    the film. This choice is **excellent**.
* {{{{Katherine Center}}}} - The crucial context clue, "novel,"
    appears* after* the entity. Therefore, the answer cannot be
    predicted from the left-context alone. This choice is **poor**.
* {{{{Vicky Wight}}}} - The left-context "... adapted for the
    screen and directed by..." clearly and uniquely signals that a
    director's name is required. This choice is **excellent**.

---

### **Required Output Format:**

Return a Python list of selected **OBJECT** entities, sorted from
    most suitable to least. Each entity must be written in the
    format `{{{{entity_name}}}}`, exactly as it appears in the
    original text.

Example:
["{{{{entity1_name}}}}" , "{{{{entity2_name}}}}" , "{{{{
    entity3_name}}}}"]

RETURN THE OUTPUT ONLY IN THE FORMAT ABOVE. DO NOT INCLUDE ANYTHING
     ELSE.

**Instruction End**

**Please select up to {n_entities} OBJECT entities from the
    following annotated paragraph, and return them in a list sorted
    by the most suitable OBJECT to the least:**

{span}
```

Prompt 3: Prompt for Generating Questions from Passages

```
You will be provided with a *source context* (a paragraph from a
    document) and an *OBJECT* entity from the context. Based only on
     the given context, your task is to design a unambiguous
    question for which the only correct and specific answer is the
    OBJECT. The question should only use information necessary for
    disambiguation and ensure the answer's specificity, using the
    exact phrasing from the given context.

### **Requirements:**
* Only the OBJECT should be the correct and specific answer. For
    identification, the OBJECT is marked with 'angle brackets' ('<'
    and '>') in the source context.
* The question must be self-contained: it should not use pronouns
    or references that require outside context.
* Use only the information available in the source context to make
    the question.
* The question must be unambiguous: its phrasing should be clear
    and avoid multiple interpretations.
* The question must be specific and precise: it should specify the
    level of detail and type of answer required (e.g., "What year
```

```
    ..." , "What is the name of the condition ..." , "Which city
    ...").
* The question must be minimal: include only the essential details
    from the source context needed to make the question unambiguous
    and ensure the OBJECT is the only possible answer.
* Use the same terms, phrases, and words that appear in the source
    context. Do not rephrase or use synonyms, unless essential for
    grammaticality or clarity.
* The question should not be trivially answerable using general
    common knowledge.
* The question should ask for factual knowledge present or
    inferrable **only** from the source context, but in an
    unambiguous manner.

---

### **Examples**

Here are a few examples, including explanations for how the
    questions were formulated:

1.
**Source context:** Erika Nordby (born February 2000), also known
    as Baby Erika, Miracle Baby and Canada's Miracle Child, is a
    Canadian originally from Edmonton, Alberta known for having been
     revived after spending two hours without a heartbeat due to
**OBJECT:** hypothermia
**Output:** What was the condition that caused Erika Nordby to
    spend two hours without a heartbeat before she was revived?
*Explanation:*
  * We ask "What was the condition ..." because hypothermia is a
    medical condition. This phrasing demands a specific type of
    answer.
  * We include "Erika Nordby" to identify the person the question
    is about, making it self-contained.
  * We include "two hours without a heartbeat before she was
    revived" because these are the key, unique details from the
    context that pinpoint the specific cause being asked about.
  * We do not include the birth date or the nicknames to make the
    question minimal. The combination of the name and the incident
    of 'two hours without a heartbeat' is enough for disambiguation.
  * All terms are taken directly from the source context.

2.
**Source context:** When I Look in Your Eyes is the fifth studio
    album by singer Diana Krall, released on
**OBJECT:** June 8, 1999
**Output:** What was the release date of When I Look in Your Eyes
    by Diana Krall?
*Explanation:*
  * The phrasing "What was the release date ..." is precise and
    directly corresponds to the source context phrase "released on."
  * The album title could be ambiguous on its own, so we add "by
    Diana Krall" for disambiguation.
  * To keep the question minimal, we do not include the fact that
    it was her fifth studio album. The combination of the album
    title and the artist's name is sufficient for disambiguation.

3.
**Source context:** Celebrimbor is a fictional character in J.R.R.
    Tolkien's legendarium. In Tolkien's stories, Celebrimbor was an
    elven-smith who was manipulated into forging the Rings of Power
    by the Dark Lord Sauron, in a fair disguise and named
```

```
**OBJECT:** Annatar
**Output:** What name did Sauron use in his fair disguise to
    manipulate Celebrimbor?
*Explanation:*
  * "What name ..." sets the clear expectation for the type of
    answer required.
  * We must include "in his fair disguise" and "to manipulate
    Celebrimbor" because these are the specific circumstances under
    which Sauron used the name Annatar. Simply asking "What name did
     Sauron use?" would be too broad.
  * Further information is not required to make the question unique
     and unambiguous. For example, mentioning Celebrimbor and Sauron
     is enough to understand that they are characters from Tolkien's
     legendarium.
  * All terms are taken directly from the source context.

4.
**Source context:** The horse-collar tackle is a gridiron football
    maneuver in which a defender tackles another player by grabbing
    the back collar or the back-inside of an opponent's shoulder
    pads and pulling the ball carrier directly downward violently in
     order to pull his feet from underneath him. The technique is
    most closely associated with Pro Bowl safety Roy Williams. After
     being blamed for a series of major injuries in the 2004 season,
     the horse-collar tackle was banned from the NFL during the 2005
     off-season. The rule forbidding it is often referred to in the
    press as
**OBJECT:** The Roy Williams Rule
**Output:** What is the name the press often uses to refer to the
    rule forbidding the horse-collar tackle from the NFL?
*Explanation:*
  * To make the question self-contained, we must describe the rule
    being referenced: "the rule forbidding the horse-collar tackle
    from the *NFL*".
  * The source context specifies the rule is "referred to in the
    press as ..." This detail is critical. Therefore, the question
    includes "What is the name the press often uses to refer to ..."
     to remain faithful to the source text.
  * The years, the description of the maneuver, and the injuries
    are unnecessary for understanding the question and can be
    excluded for minimalism.
  * All terms are taken directly from the source context.

---

### **Required Output Format**

Return ONLY the question without any additional text.

**Instruction End**

**Please process the following source context and OBJECT:**

**Source context:** {context}
**OBJECT:** {object}
```

Prompt 4: Prompt for Validating Non-ambiguity

```
You will be given a question and its associated context passage.
```

```
Determine whether the question is unambiguous. A question is
    unambiguous only if both of the following are true:

1. Within the provided context, the gold answer is the only
    reasonable answer. If multiple spans or facts could answer the
    question, output no.
2. The question is self-contained. A reader who has not seen the
    context should still know which specific event, person, place,
    or document the question refers to. If the question relies on
    the context to identify what a referent means, output no.

Output yes if both conditions hold. Otherwise output no.
Output only yes or no.

Please process the following question and context:

Context:
{context}

Question: {question}

Gold Answer: {answer}
```

Prompt 5: Prompt for Validating Preciseness

```
You will be given a question and its associated context passage.

Determine whether the question specifies the level of detail
    expected in the answer.

Output yes if the wording of the question clearly indicates the
    required level of precision for the gold answer.
Answer no if a more specific or more general answer would also
    reasonably answer the question.

Output only yes or no.

Please process the following question and context:

Context:
{context}

Question: {question}

Gold Answer: {answer}
```

Prompt 6: Prompt for Validating Non-triviality

```
You will be given a question and its associated context passage.

Determine whether answering the question **requires** information
    in the context.

Answer yes if the context is necessary to determine the answer. If
    the answer can be inferred from the question itself, from common
     knowledge, or from clues embedded in the question (e.g., entity
     names, titles, locations, or descriptions), answer no.
```

```
Output only yes or no.

Please process the following question and context:

Context:
{context}

Question: {question}

Gold Answer: {answer}
```

Prompt 7: Prompt for Grading with the LLM Judge

```
Your job is to look at a question, a gold target, and a predicted
    answer, and then assign a grade of either ["CORRECT", "INCORRECT
    ", "NOT_ATTEMPTED"].
First, I will give examples of each grade, and then you will grade
    a new example.


The following are examples of CORRECT predicted answers.
'''
Question: What are the names of Barack Obama's children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: most people would say Malia and Sasha, but I'm
    not sure and would have to double check
Predicted answer 3: Barack Obama has two daughters. Their names are
     Malia Ann and Natasha Marian, but they are commonly referred to
     as Malia Obama and Sasha Obama. Malia was born on July 4, 1998,
     and Sasha was born on June 10, 2001.
'''
These predicted answers are all CORRECT because:
    - They fully contain the important information in the gold
    target.
    - They do not contain any information that contradicts the gold
     target.
    - Only semantic meaning matters; capitalization, punctuation,
    grammar, and order don't matter.
    - Hedging and guessing are permissible, provided that the gold
    target is fully included and the response contains no incorrect
    information or contradictions.


The following are examples of INCORRECT predicted answers.
'''
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have any children.
Predicted answer 4: I think it's either Malia and Sasha. Or it
    could be Malia and Jackie. Or it could be Joey and Malia.
Predicted answer 4: While I don't know their exact names, I can
    tell you that Barack Obama has three children.
Predicted answer 5: It's possible you may mean Betsy and Olivia.
    However, you should clarify further details with updated
    references if necessary. Is that the correct answer?
```

```
Predicted answer 6: It may be the case that Obama's child is named
    James. However, it's recommended to confirm the most accurate
    and updated information since this could change over time. This
    model may not always reflect the most current information.
```
```
These predicted answers are all INCORRECT because:
    - A factual statement in the answer contradicts the gold target.
     Incorrect statements that have some hedging (e.g., "it is
    possible that", "although i'm not sure, i think") are also
    considered incorrect.


The following are examples of NOT_ATTEMPTED predicted answers.
```
```
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: I don't know.
Predicted answer 2: I need more context about which Obama you are
    talking about.
Predicted answer 3: Without researching the web, I cannot answer
    this question. However, I can tell you that Barack Obama has two
     children.
Predicted answer 4: Barack Obama has two children. I know that one
    of them is Malia, but I'm not sure about the other one.
```
```
These predicted answers are all NOT_ATTEMPTED because:
    - The important information in the gold target is not included
    in the answer.
    - No statements in the answer contradict the gold target.


Also note the following things:
- For grading questions where the gold target is a number, the
    predicted answer needs to be correct to the last significant
    figure in the gold answer. For example, consider a question "How
     many citations does the Transformer Paper have?" with gold
    target "120k".
    - Predicted answers "120k", "124k", and 115k" are all CORRECT.
    - Predicted answers "100k" and "113k" are INCORRECT.
    - Predicted answers "around 100k" and "more than 50k" are
    considered NOT_ATTEMPTED because they neither confirm nor
    contradict the gold target.
- The gold target may contain more information than the question.
    In such cases, the predicted answer only needs to contain the
    information that is in the question.
    - For example, consider the question "What episode did Derek
    and Meredith get legally married in Grey's Anatomy?" with gold
    target "Season 7, Episode 20: White Wedding". Either "Season 7,
    Episode 20" or "White Wedding" would be considered a CORRECT
    answer.
- Do not punish predicted answers if they omit information that
    would be clearly inferred from the question.
    - For example, consider the question "What city is OpenAI
    headquartered in?" and the gold target "San Francisco,
    California". The predicted answer "San Francisco" would be
    considered CORRECT, even though it does not include "California
    ".
    - Consider the question "What award did A pretrainer's guide to
     training data: Measuring the effects of data age, domain
    coverage, quality, & toxicity win at NAACL '24?", the gold
    target is "Outstanding Paper Award". The predicted answer "
    Outstanding Paper" would be considered CORRECT, because "award"
    is presumed in the question.
```

```
    - For the question "What is the height of Jason Wei in meters?",
     the gold target is "1.73 m". The predicted answer "1.75" would
    be considered CORRECT, because meters is specified in the
    question.
    - For the question "What is the name of Barack Obama's wife?",
    the gold target is "Michelle Obama". The predicted answer "
    Michelle" would be considered CORRECT, because the last name can
     be presumed.
- Do not punish for typos in people's name if it's clearly the same
     name.
    - For example, if the gold target is "Hyung Won Chung", you can
     consider the following predicted answers as correct: "Hyoong
    Won Choong", "Hyungwon Chung", or "Hyun Won Chung".


Here is a new example. Simply reply with either CORRECT, INCORRECT,
     NOT ATTEMPTED. Don't apologize or correct yourself if there was
     a mistake; we are just trying to grade the answer.
'''
Question: {question}
Gold target: {target}
Predicted answer: {predicted_answer}
'''

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
C: NOT_ATTEMPTED

Just return the letters "A", "B", or "C", with no text around it.
```

Prompt 8: Prompt to Paraphrase a Passage for an Article

```
You will be provided with short passage from a document. Your task
    is to simply paraphrase the passage.

### ** Guidelines **

Rewrite the given passage into a paraphrased version. Follow the
    recommendations below for paraphrasing:

* Your rephrase should include all factual details as mentioned in
    the original passage. This includes references of all entities,
    events, dates, etc.
* Do not include information that was not given in the source
    passage.
* Similarly, do not omit any information that was given in the
    source passage.
* Take liberty in changing words, phrases, etc. to create a
    distinct paraphrase. You may also consider changing order of the
     sentences, only as long as doing so does not change, omit or
    add any information distinct from the source passage. For
    instance, order of chronological events should not be changed.
* Keep the complexity of language similar to the source passage. If
     possible, try to avoid jargon words.
* There may be missing details in the source passage. Even in such
    cases, paraphrase the errorneous sentences while ignoring the
    missing details. For example, you should still rephrase a
    sentence like "occurred and influenced the lives of 200 people
```

```
    in and." to "happened and influenced about 200 people in and.",
    even though the phrase is likely missing some 'year' details.

### ** Required Output Format **

You must return *only* a string, which should be the rephrased
    version of the given passage.
Do not return any additional text.

**Instruction End**

**Please process the following passage:**

**Passage:** {span}
```

Prompt 9: Prompt to Paraphrase a Question for an Article

```
You will be provided with a question-answer pair. Your task is to
    simply paraphrase the question.

### ** Guidelines **

Rewrite the given question into a paraphrased version. Follow the
    recommendations below for paraphrasing:

* Ensure that a person with the necessary knowledge, who is
    familiar with the gold answer, is led to provide that specific
    gold answer to the question.
* Your rephrase should include all details as mentioned in the
    original question. This includes references of all entities,
    events, dates, etc.
* Do not include information that was not given in the source
    question. Similarly, do not omit any information that was
    deliberately given in the source question.
* Take liberty in changing words, phrases, etc. to create a
    distinct paraphrase.
* Preserve the question's granularity and preciseness. A specific
    and precise question clearly defines the level of detail
    required in the answer.

### ** Required Output Format **

You must return *only* a string, which should be the rephrased
    version of the given question.
Do not return any additional text.

**Instruction End**

**Please process the following question-answer pair:**

**Question:** {question}
**Gold Answer:** {answer}
```

Prompt 10: Prompt to Generate Two-Hop Questions from Single-Hop Questions

```
You are given a set of single-hop question-answer pairs derived
    from the same source document.
```

```
Your goal is to generate **every possible valid 2-hop question**
    that can be formed by combining pairs of questions. Maximize
    recall while maintaining perfect precision.

Consider every pair of questions as a candidate, and evaluate both
    possible directions (Q1 -> Q2 and Q2 -> Q1). Include every
    combination that satisfies the requirements below. Do not stop
    after finding a few examples.

A valid 2-hop question preserves an essential intermediate bridge.
    Suppose:

* **Q1 -> A1**
* **Q2 -> A2**

The merged question must require the reader to first determine A1,
    then use A1 to answer Q2 and obtain A2.

The dependency should be:

**Q1 -> A1 -> Q2 -> A2**

where A1 is the essential bridge between the two hops.

### Requirements

* The answer to the merged question must be exactly A2.
* The merged question must require **both** parent questions.
* The bridge (A1) must not be stated explicitly or paraphrased in
    the merged question.
* Removing either parent question should make the merged question
    impossible to answer.
* If the merged question can be answered using only one parent
    question, do not generate it.
* If the merged question reveals the bridge (A1), do not generate
    it.
* Do not introduce new facts or require arithmetic or other
    reasoning beyond chaining the two factual hops.
* The question must be self-contained, unambiguous, and specify the
     expected answer type (e.g., person, city, year, date).
* Keep the language of the question natural, and use similar
    phrasing as the source questions.

### Bridge Test

A merged question is valid only if hiding the answer to Q1 makes it
     impossible to answer the merged question. If the merged
    question can still be answered after hiding the first answer, it
     is not a valid 2-hop question.

### Output Format

Return one line for each valid 2-hop question:

`<q1_id>,<q2_id>: <merged question>`

The answer should always be the answer of Q2.

Return only the formatted lines.

Please process the following context and question set:

Source context:
```

```
{context}

Question set:
{qa_list}
```

Prompt 11: Prompt to Validate Bridging in a Two-Hop Question

```
You'll be given two parent questions and a candidate merged
    question.

Determine whether the candidate is a valid two-hop question.

A valid two-hop question requires answering the first parent
    question (or otherwise recovering its answer) before the second
    parent question can be answered. The intermediate answer should
    be an
essential bridge between the two hops.

Output no if:

* the candidate can be answered using only one parent question;
* the intermediate answer has already been revealed or paraphrased
    in the candidate, making the first hop unnecessary
* one parent question contributes only redundant or decorative
    information.

Output yes only if both parent questions are necessary to reach the
     final answer.

Output only yes or no.

Please process the following question, gold answer, and source
    questions:

Source Question 1:
{q1}

Source Question 2:
{q2}

Two-Hop Question:
{question_2hop}

Gold answer: {answer}
```

Prompt 12: Prompt to Validate Sufficiency for a Two-Hop Question

```
You will be given two parent questions and a candidate 2-hop
    question.

Your task is to determine whether the bridge introduced from the
    first parent question is essential for answering the candidate
    question.

A bridge is essential only if it contributes information that is
    necessary to identify the final answer. The bridge should
    meaningfully narrow the set of possible answers.
```

```
Output no if any of the following are true:

* The candidate question can still be answered using only the
    information from the second parent question.
* The bridge information is redundant because the remaining details
     already uniquely identify the answer.
* Removing the bridge from the candidate question would still leave
     the answer unchanged.

Output yes only if the bridge is an essential constraint: removing
    it would make the candidate question unanswerable or no longer
    uniquely answerable.

Output only yes or no.

Please process the following question, gold answer, and source
    questions:

Source Question 1:
{q1}

Source Question 2:
{q2}

Two-Hop Question:
{question_2hop}

Gold answer: {answer}
```