The Role of Mixed-Language Documents for Multilingual Large Language Model Pretraining

Jiandong Shao,[2] **Raphael Tang,**[1] **Crystina Zhang,**[3] **Karin Sevegnani,**[4]
Pontus Stenetorp,[1,5] **Jianfei Yang,**[2] **Yao Lu**[1*]

[1]University College London [2]Nanyang Technological University
[3]University of Waterloo [4]NVIDIA [5]National Institute of Informatics

Abstract

Multilingual large language models achieve impressive cross-lingual performance despite largely monolingual pretraining. While bilingual data in pretraining corpora is widely believed to enable these abilities, details of its contributions remain unclear. We investigate this question by pretraining models from scratch under controlled conditions, comparing the standard web corpus with a monolingual-only version that removes all multilingual documents. Despite constituting only 2% of the corpus, removing bilingual data causes translation performance to drop 56% in BLEU, while behaviour on cross-lingual QA and general reasoning tasks remains stable, with training curves largely overlapping the baseline. To understand this asymmetry, we categorize bilingual data into parallel (14%), code-switching (72%), and miscellaneous documents (14%) based on the semantic relevance of content in different languages. We then conduct granular ablations by reintroducing parallel or code-switching data into the monolingual-only corpus. Our experiments reveal that parallel data almost fully restores translation performance (91% of the unfiltered baseline), whereas code-switching contributes minimally. Other cross-lingual tasks remain largely unaffected by either type. These findings reveal that translation critically depends on systematic token-level alignments from parallel data, whereas cross-lingual understanding and reasoning appear to be achievable even without bilingual data. To promote reproducibility and facilitate further research, we release our corpus and models under Open Source Initiative approved licenses.[1]

1 Introduction

Large language models (LLMs), when pretrained on web-collected data from different language sources, exhibit remarkable emergent capabilities in cross-lingual understanding despite not being pretrained using multilingual-specific objectives (Devlin et al., 2019; Achiam et al., 2023; Yang et al., 2024). Existing research attributes this behaviour not only to sufficient data from different languages, but also to specific documents where multiple languages co-occur in the same context (Chaudhary et al., 2020; Chi et al., 2020; Wang et al., 2025). Motivated by this observation, multilingual pretraining strategies often incorporate multilingual data, under the hypothesis that mixed-language exposure uniformly benefits cross-lingual tasks (Yoo et al., 2024; Wang et al., 2025).

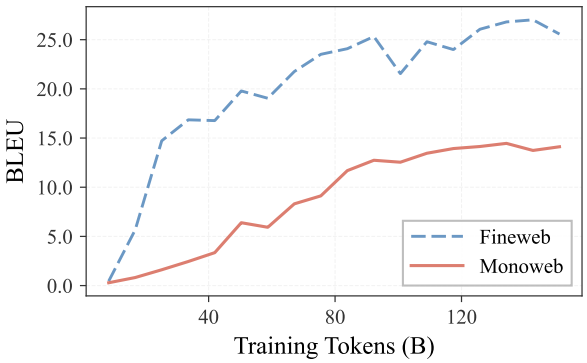

Figure 1: Performance on WMT14 for different pretraining setups. FINEWEB: multilingual web data from FineWeb and FineWeb2; MONOWEB: multilingual web data with bilingual documents *removed*.

However, the high cost of pretraining and large-scale pretraining data classification has constrained the scope of existing explorations of the role of multilingual data. Studies that rely on continual pretraining typically build on models that may have already been exposed to related data during pretraining, which makes the role of multilingual data more difficult to disentangle. Among the few works that investigate multilingual data at the pretraining stage, existing studies (Briakou et al., 2023; Qorib et al., 2025; Wang et al., 2025) do not provide a

*Corresponding author.

[1]Dataset: https://hf.co/datasets/UCLNLP/monoweb-dataset; Models: https://hf.co/UCLNLP/monoweb

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36807–36818

Category	Example
Parallel	Magnifique et lumineux loft Toronto de 1 chambre avec plafonds de 10 pi et grande terrasse extérieure comprenant un barbecue.... Beautiful, bright one bedroom Toronto loft with 10ft ceilings and large outdoor terrace including barbeque..... *[Paragraph-aligned translation with systematic cross-lingual correspondence]*
Code-switching	The people, filled with joy, chant the anthem "A qua ben fé! A qua ben fé! La tarascou a rou un bré!" *[Natural language mixing within shared discourse context]*
Miscellaneous	...and in some cases whether to let the fires burn to create regeneration in the forest. Vous devez avoir la dernière version de Flash Player installée. *[French text about Flash Player, semantically unrelated to the English content]*

Table 1: Examples of MONOWEB filtered data. We classify documents into three categories: documents with clear parallel structure, documents that exhibit code-switch behaviour, and miscellaneous documents.

systematic analysis of its role, but instead focus on specific settings or mechanisms. To this end, we aim to conduct a thorough analysis of the pretraining corpus and design a controlled pretraining setup to reveal the role of multilingual data.

We construct a monolingual web corpus by filtering out all documents containing more than one language from standard web-collected data. This procedure removes fewer than 2% of documents, making fine-grained analysis of multilingual data feasible at scale. We then pretrain multilingual LLMs from scratch under two setups: MONOWEB, using the filtered corpus, and FINEWEB, using the original web data. Despite accounting for only 2% of pretraining data, multilingual documents are critical for machine translation. Removing them causes BLEU scores to drop from 22.3 to 9.8 (a 56% relative decrease), effectively collapsing translation performance. In contrast, other cross-lingual tasks are substantially less affected: cross-lingual QA drops by 10% on average, while understanding and reasoning tasks vary by at most 4%. This asymmetry highlights the nuanced role of pretraining data across different multilingual tasks.

To better understand this phenomenon, we analyze the composition of the removed multilingual documents. We find that most consist of bilingual content, which can be grouped into three categories (Table 1): (i) *parallel documents* (14%), providing aligned translations with explicit cross-lingual correspondence, such as multilingual Airbnb webpages; (ii) *code-switching documents* (72%), where languages naturally alternate within shared discourse, as commonly observed in user-generated content on platforms like Pinterest; and (iii) *miscellaneous documents* (14%), where multiple languages co-occur without meaningful semantic alignment. We then isolate the contributions of different bilingual data types through controlled pretraining from scratch. Our results show that parallel data, despite comprising only 14% of bilingual documents, is the dominant factor for translation performance: reintroducing only parallel data yields a 106% improvement over MONOWEB, largely recovering performance relative to FINEWEB (BLEU 20.2 vs. 22.3). In contrast, reintroducing code-switching data provides only marginal gains (BLEU 12.4 vs. 9.8), with little effect on other cross-lingual tasks. Finally, we analyze the underlying causes of this asymmetry. We find that removing bilingual data primarily disrupts lexical-level cross-lingual alignment, leading to severe translation failures, while sentence-level alignment remains largely preserved, explaining the robustness of non-translation tasks.

To summarise, our contributions are threefold:

1. We introduce a monolingual dataset, MONOWEB, together with a detailed analysis of multilingual content, pretrain models from scratch to study multilingual behavior without mixed-language exposure, and open-source both the dataset and models.

2. Through pretraining from scratch, we demonstrate a task-dependent sensitivity to bilingual data: machine translation critically depends on a tiny fraction (less than 2%) of bilingual documents, whereas other cross-lingual understanding and reasoning tasks remain largely unaffected. We further show that different types of bilingual data contribute unequally, with parallel data playing a disproportionately critical role.

3. We provide an in-depth failure mode and representation-level analysis, revealing that

the degradation in translation performance is driven by the loss of lexical-level alignment, while sentence-level alignment remains largely preserved.

2 Related Work

Multilingual data is widely assumed to drive cross-lingual capabilities in multilingual models. Parallel corpora (sentence-aligned translations) are well-known to be essential for machine translation (Brown et al., 1993), enabling multilingual MT systems to bridge high- and low-resource languages (Johnson et al., 2016; Fan et al., 2020; team et al., 2022; Aycock et al., 2024). Beyond parallel data, naturally occurring code-switching, where languages alternate within the same discourse, has also attracted attention as a potential mechanism for cross-lingual alignment. Prior work demonstrates that using code-switching for data augmentation can improve zero-shot transfer during finetuning (Qin et al., 2020), and that curriculum learning with code-switching enhances transfer to low-resource languages (Yoo et al., 2024). These results have motivated practitioners to incorporate multilingual content under the assumption that mixed-language exposure benefits cross-lingual tasks (Qorib et al., 2025).

However, most existing studies focus on finetuning or continued training, which remains limited because models may have already been exposed to similar data during the pretraining stage. Among the few works that investigate multilingual data in the pretraining stage, most focus on specific aspects rather than systematically studying its role: one primarily characterizes incidental bilingualism in existing corpora (Briakou et al., 2023), another uses generated data to study curriculum learning effects (Qorib et al., 2025), and a third explores synthetic code-switching for cross-lingual transfer (Wang et al., 2025). This leaves a gap in the understanding of the role of multilingual data during pretraining, particularly regarding the differential contributions of parallel versus code-switching data. We aim to fill this gap by systematically ablating different bilingual data types and studying their impact on multilingual LLMs.

3 MonoWeb Pretraining Data

To study the heterogeneous role of bilingual data, we first construct a multilingual corpus by sampling 60B tokens per language from FineWeb-Edu (Lozhkov et al., 2024) (English) and FineWeb2 (Penedo et al., 2025) (German, Spanish, French), totaling 240B tokens. We then perform a systematic characterization of the bilingual documents in this corpus, focusing on English-paired bilingual content (en-de, en-es, en-fr) as English serves as the current dominant lingua franca for cross-lingual scenarios.

3.1 Bilingual Data Identification

We identify bilingual documents through a two-stage pipeline, combining rule-based filtering with LLM-based classification to ensure both scalability and accuracy.

Stage 1: Candidate Detection via Entropy-based Filtering. We first detect candidate bilingual documents using language-level entropy as a proxy for language mixing. For each document, we perform sentence segmentation using NLTK (Bird et al., 2009) and apply fastText language identification (Joulin et al., 2016) to compute language confidence scores for each sentence. Taking English-French as an example, for each sentence s_i with length l_i, fastText outputs confidence scores for English (p_i^{en}) and French (p_i^{fr}). We then compute a document-level language distribution by aggregating sentence-level confidence scores weighted by sentence length:

$$P_{\text{doc}}^{\text{lang}} = \frac{\sum_i l_i \cdot p_i^{\text{lang}}}{\sum_i l_i}, \tag{1}$$

where lang $\in \{\text{en}, \text{fr}\}$. After normalization, we obtain a probability distribution over the two languages for the entire document. We compute the entropy of this distribution:

$$H = -\sum_{\text{lang}} P_{\text{doc}}^{\text{lang}} \log P_{\text{doc}}^{\text{lang}}. \tag{2}$$

Documents with entropy above a threshold $\tau = 0.1$ (indicating substantial mixing of both languages) are marked as bilingual candidates. We empirically selected this threshold by examining the distribution of entropy values and verifying that it effectively captures documents with substantial language mixing. This stage serves as a coarse filtering that optimizes for the recall of potential bilingual data, while maintaining computational efficiency. As a result, 5% of the corpus is retained and can be processed using more computationally expensive methods during the subsequent verification stage.

Stage 2: LLM-based Classification. To distinguish different types of bilingual relationships, we employ LLAMA-3.3-70B-INSTRUCT (Dubey et al., 2024) for a *two-step classification* process, whose reliability has been validated through human evaluation. First, the model verifies whether each candidate is genuinely bilingual, which aims to filter out the false positives included by entropy-based filtering. We consider the resulting set of documents after this step as the final verified bilingual documents, which consists of approximately 2% of the entire corpus. Second, based on the semantic relationship of contents in different languages, the verified bilingual documents are classified into one of the three categories:

- **Parallel documents**: Paragraph-aligned translations where languages express identical semantic content with systematic correspondences (e.g., dictionaries, translated website; the example in Table 1 is from Airbnb).
- **Code-switching documents**: Documents where both languages appear with semantic relationships but without systematic alignment. This includes naturally occurring mixed-language discourse (e.g., multilingual forum discussions), articles with embedded foreign quotations or terminology, and documents where languages serve complementary communicative functions. Crucially, unlike parallel data, the two languages do not provide translations of each other but rather contribute distinct yet related semantic content.
- **Miscellaneous documents**: Documents where multiple languages co-occur without meaningful cross-lingual semantic relationships. This category primarily consists of web artifacts such as multilingual boilerplate, advertisements in different languages, or navigation elements appended to otherwise monolingual content.

This two-stage approach balances computational efficiency with classification accuracy: entropy-based filtering reduces the search space from the full corpus to 5% candidates, while LLM classification provides semantic nuance that rule-based methods lack. Table 1 shows representative examples for each category. The resulting taxonomy enables granular ablations to isolate the effects of different bilingual data types during pretraining.

Human Validation. To assess the reliability of the classification, we manually inspected 50 randomly sampled documents from each category (150 in total). The results show an overall agreement rate of 90% between human annotations and the LLM-based classifier. Discrepancies are primarily observed in exceptionally long documents, likely due to the limitations of the LLM context window. Given that bilingual documents account for less than 2% of the entire corpus, such classification noise is expected to have a negligible impact on the overall conclusions.

3.2 Bilingual Data Analysis

Table 2 presents the statistics of bilingual data in our corpus, where the pattern is similar for all three language pairs: the bilingual data constitutes 2% of the entire 240B-token pretraining corpus, and it is dominated by code-switching data ($> 70\%$), and a similar amount of parallel and miscellaneous documents (10-20%). We further analyze the website URLs of the parallel and code-switching documents to understand the main sources of each categories of bilingual data, reported in Table 3.

Parallel data, while comprising less than 20% of bilingual data, originates from high-quality curated sources. As Table 3 reveals, academic repositories dominate, particularly doctoral theses with multilingual abstracts. Bilingual dictionaries and language learning platforms (reverso.net) provide sentence-aligned translations, while technical documentation (docs.microsoft.com) contributes systematic correspondences. These sources feature explicit token-level alignments where each segment has an equivalent in another language.

Code-switching dominates at 72% of bilingual data, which originates primarily from social content aggregation sites (pinterest.com), E-commerce platforms with mixed-language reviews (amazon.fr), and forums.

The remaining miscellaneous 14% consists of noise—multilingual boilerplate and web artifacts where languages accidentally co-occur without meaningful relationships.

Overall, the URL analysis reveals a clear fundamental distinction on the source of bilingual data under different categories: parallel data provides professionally curated alignments from dictionaries and academic repositories, while code-switching reflects spontaneous language mixing in user-generated content.

Data Type	en-de	en-es	en-fr
Bilingual data in Corpus			
Total Bilingual	2.80%	1.62%	2.40%
Bilingual Data Composition			
Parallel	10%	17%	15%
Code-switching	75%	69%	73%
Miscellaneous	15%	14%	12%

Table 2: Bilingual data statistics for each language pair. The top section reports the proportion of bilingual data in the full corpus, showing that such data is generally sparse; the bottom section shows the distribution of the bilingual data types.

Type	Representative Sources	%
Parallel	Academic (thesis.fr)	35
	Dictionaries (reverso.net)	15
	Travel (airbnb.com)	15
	Canadian (umontreal.ca)	6
	Professional (docs.microsoft)	8
Code-switching	Social (pinterest.com)	25
	Forums (forumactif.com)	10
	E-commerce (amazon.fr)	8

Table 3: Approximate domain distribution of bilingual data based on URL analysis of the top 50 sources from the en-fr corpora. Parallel data originates mainly from academic sources and dictionaries with systematic alignments, while code-switching appears in user-generated content with organic language mixing.

4 Experimental Setup

4.1 Pretraining Configurations

We conduct experiments on three language pairs: English-French (en-fr), English-German (en-de), and English-Spanish (en-es). For each language pair, we construct a bilingual corpus by combining 60B English tokens with 60B tokens from the target language (French, German, or Spanish), sampled from FineWeb-Edu and FineWeb2.

For each language pair, we pretrain models using four data configurations:

- **FINEWEB**: Full 120B-token corpus including all bilingual data.
- **MONOWEB**: All bilingual documents removed, retaining only monolingual content.
- **MONOWEB+PARALLEL**: MONOWEB augmented with only parallel documents.
- **MONOWEB+CODESWITCH**: MONOWEB augmented with only code-switching documents.

We exclude miscellaneous data as it lacks cross-lingual semantic relationships. This yields 12 models in total (3 language pairs × 4 configurations), all trained from scratch to ensure a fair comparison.

4.2 Model Architecture and Training

We train decoder-only transformer models with 1.35B parameters using the Llama-2 tokenizer (Touvron et al., 2023) (32K vocabulary). The architecture consists of 24 layers with a hidden dimension of 2048, 16 attention heads, and a 2048 context length. All models are trained for 34K steps (1̃43B tokens) with a batch size of 2,048 using the AdamW optimizer (Loshchilov and Hutter, 2017) and a 6e-4 learning rate, including 2,000 warmup steps followed by constant decay. We set weight decay to 0.1, apply gradient clipping at 1.0, and use Adam betas of 0.9 and 0.95. Training is performed with Megatron-LM (Shoeybi et al., 2019) and takes about 6,144 A100 GPU hours per model.

4.3 Downstream Evaluation Suite

All tasks are evaluated using the lm-evaluation-harness (Gao et al., 2024), along with five-shot prompting and default configurations.

Machine Translation We evaluate translation quality on standard benchmarks for all three language pairs: wmt16 en-de (Bojar et al., 2016), wmt14 en-fr (Bojar et al., 2014), and flores-101 en-es (Goyal et al., 2021), testing both translation directions and report BLEU scores (Papineni et al., 2002) separately for each direction.

Cross-lingual Question Answering We evaluate cross-lingual question answering using two complementary benchmarks. (1) For XQuAD (Artetxe et al., 2019), we adapt the dataset by placing the context in language L1 and both the question and answer in language L2, allowing us to assess the model's ability to generate answers across languages. (2) For MLQA (Lewis et al., 2019), we follow the original setup, where the context and answer are in language L1 and the question is in language L2, which primarily evaluates the model's ability to retrieve information across languages. We report Exact Match scores for all language pairs.

Cross-lingual Understanding and Reasoning We evaluate models on a suite of benchmarks covering both cross-lingual understanding and reasoning

abilities. For cross-lingual natural language understanding, we use XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019) to assess whether bilingual data improves the transfer of inference and paraphrase recognition skills. For reasoning tasks, we include HellaSwag (Zellers et al., 2019; Dac Lai et al., 2023) for commonsense reasoning, ARC (Dac Lai et al., 2023; Clark et al., 2018) for knowledge-intensive reasoning, TruthfulQA (Lin et al., 2022; Dac Lai et al., 2023) for factual consistency, and additionally XStoryCloze (Lin et al., 2021) (en, es) and XWinograd (Tikhonov and Ryabinin, 2021) (en, fr) for narrative comprehension and coreference resolution. We report accuracy for all tasks.

5 Results

Tables 5, 7, and 8 present results across all tasks and configurations. A clear task-specific asymmetry emerges: removing bilingual data causes significant degradation for machine translation (56% BLEU drop), moderate decline in cross-lingual QA (<10%), and has almost no effect on understanding and reasoning tasks. This indicates different levels of reliance on bilingual exposure across tasks, suggesting that different cross-lingual abilities may rely on qualitatively different learning signals.

5.1 Machine Translation: Critical Dependence on Parallel Data

Table 5 summarizes translation results across all configurations and language pairs. Removing all bilingual data leads to substantial performance degradation, with average BLEU dropping from 22.3 to 9.8 (56% relative decline). Reintroducing only parallel documents, which comprise 10–17% of bilingual content, largely recovers performance (20.2 BLEU, 91% of the original performance). In contrast, adding back code-switching text (72% of the bilingual data) yields only a minimal improvement (12.4 BLEU, corresponding to 56% of the original performance).

This pattern is consistent across all six translation directions (Table 5). Individual language pairs show a 41–69% relative degradation when bilingual data is removed, and 90–107% recovery when parallel data alone is reintroduced.

These results highlight that translation quality depends critically on explicit cross-lingual alignment rather than incidental code-switching.

Test Set Contamination. The large performance gap between FINEWEB and MONOWEB on translation raises the question of whether FINEWEB benefits from test set contamination. To rule this out, following (Brown et al., 2020; Xu et al., 2025), we perform a 10-gram overlap analysis between the parallel subset of our pretraining corpus and all MT evaluation benchmarks. The overlap rate is negligible across all language pairs (En-De: 0.00%, En-Fr: 0.023%, En-Es: 0.012%), and manual inspection confirms that matches are driven by generic proper nouns rather than benchmark sentences, ruling out contamination as an explanation for the observed performance gap.

5.2 Understanding Translation Collapse: A Two-Fold Failure

To understand the mechanisms behind translation performance degradation, we analyze 1,000 sampled En→De translation outputs, using Llama-3.3-70B-Instruct as a language identifier to classify each as German, English, or mixed-language. Table 6 shows a clear disparity: FINEWEB and MONOWEB+PARALLEL generate German in more than 85% of cases, while MONOWEB and MONOWEB+CODESWITCH produce German in only around 45%. The remaining 55% are predominantly English passthroughs, which naturally yield zero BLEU contribution. These results indicate that models trained without parallel data often fail at the most basic requirement of translation—producing text in the target language.

However, language generation failure alone cannot account for the full extent of BLEU degradation. As shown in Table 6, when the evaluation is restricted to outputs that are correctly generated in German, MONOWEB still achieves only 7.70 BLEU which is less than half of FINEWEB's 17.4. Even under comparable output-language conditions, a 56% quality gap remains. MONOWEB+CODESWITCH performs even worse at 6.21 BLEU. This reveals two compounding failure modes: (1) 56% failure to generate target language (43.6% vs. 86.6% German generation), and (2) severely degraded translation quality for the remaining outputs (7.70 vs. 17.4 BLEU). The overall performance decrease compounds both problems.

To further examine the nature of the degraded translation fidelity, we manually analyzed 100 correctly generated German outputs. The analysis reveals a consistent pattern of semantic underspecification: MONOWEB captures only coarse-

Source (en)	FINEWEB (de)	MONOWEB (de)
The students should receive a grant immediately.	Die Schüler sollten **sofort** einen Zuschuss erhalten.	Die Studierenden sollten eine Unterstützung erhalten.
This was a conscious decision - diversity is an important topic here.	Dies war eine bewusste Entscheidung - **Vielfalt ist ein wichtiges Thema hier**.	Das war ein bewusster Entschluss.
He's a hero to his kids and his wife.	Er ist ein Held für seine **Kinder** und seine Frau.	Er ist ein Held für seine Familie und seine Frau.

Table 4: Fine-grained information loss in MONOWEB translations. Core propositions are preserved, but precise details are systematically lost: temporal specifications (example 1: "immediately"), explanatory contexts (example 2: diversity rationale), and lexical precision (example 3: "kids" → "Familie" [family]). Bold text shows precise translations; underlined text indicates lost information.

Direction	FWB	MWB	MWB+P	MWB+CS
en→de	16.2	5.0	17.0	4.6
de→en	24.6	14.5	21.3	14.9
en→es	17.7	6.6	17.3	11.4
es→en	21.4	8.3	20.1	16.0
en→fr	25.4	12.1	22.7	17.4
fr→en	28.6	15.3	28.8	18.7
Average	**22.3**	**9.8**	**20.2**	**12.4**

Table 5: BLEU scores for each translation direction. Removing bilingual data (MWB) causes a substantial drop, while adding parallel data (MWB+P) largely restores performance. FWB = FINEWEB, MWB = MONOWEB, P = Parallel, CS = Code-Switch.

	On-target (%)	BLEU
FWB	86.6	17.4 / 19.2†
MWB+P	89.7	17.8
MWB	43.6	7.70
MWB+CS	45.2	6.21

Table 6: On-target generation rate and translation quality on En→De. MWB and MWB+CS exhibit two compounding failures: low on-target generation rate (∼45% vs. ∼87%) and degraded translation quality on on-target outputs (7.70 vs. 17.4 BLEU). †FWB re-evaluated on the matched subset where MWB generated German; the quality gap persists (19.2 vs. 7.70), ruling out potential selection bias.

grained semantic information, preserving the basic propositional structure (who does what) but loses fine-grained information about how, when, why, and to what degree. As illustrated in Table 4, FINEWEB accurately preserves temporal and explanatory details ("immediately" → "sofort"; "diversity is an important topic here" → "Vielfalt ist hier ein wichtiges Thema"), whereas MONOWEB tends to produce paraphrases that erase such distinctions. Lexical precision also deteriorates, e.g., translating "kids" as "Familie" [family] instead of "Kinder" [kids].

Task	FWB	MWB	MWB+P	MWB+CS
German				
XQuAD	28.9	25.2	31.2	29.0
MLQA	20.6	22.4	21.4	19.1
Spanish				
XQuAD	31.8	29.7	32.1	29.9
MLQA	22.7	23.9	22.8	20.3
XQuAD Avg	**30.4**	**27.5**	**31.7**	**29.5**
MLQA Avg	**21.7**	**23.2**	**22.1**	**19.7**

Table 7: Cross-lingual QA performance averaged over both directions per language pair. XQuAD shows moderate sensitivity to bilingual data, while MLQA remains largely stable.

These observations suggest that without parallel supervision, models internalize only approximate cross-lingual alignment, resulting in content-preserving yet information-thinned translations.

5.3 Other Cross-lingual Tasks: Minimal Dependence on Bilingual Data

Cross-lingual Question Answering Table 7 presents results for XQuAD and MLQA. The two tasks show different sensitivity to bilingual data removal. For XQuAD, MONOWEB underperforms FINEWEB during training (Figure 2a), achieving 27.5 EM compared to FINEWEB's 30.4 EM (9.5% drop). On MLQA, the training curves (Figure 2b) show overlapping trajectories across configurations, with the final scores ranging from 21.7 to 23.2 EM. Unlike XQuAD, MLQA exhibits no consistent separation between configurations during training.

This difference may reflect distinct task structures: XQuAD requires generating L2 answers from L1 contexts, while MLQA primarily involves retrieving answers within L1 after understanding L2 questions. The former demands active cross-lingual generation capability, which, like machine translation, benefits from the explicit lexical-level

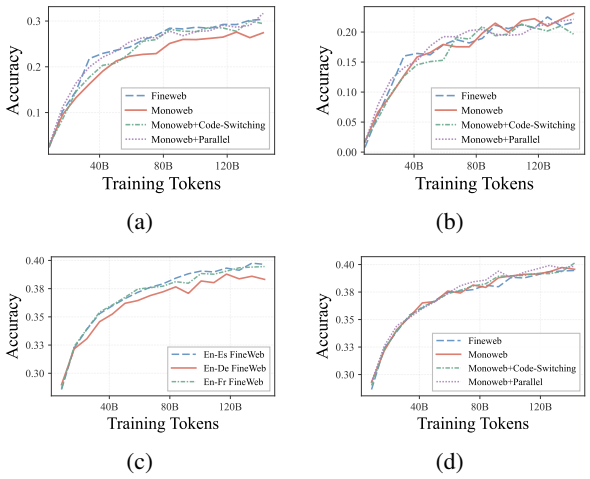

Figure 2: Training performance across cross-lingual tasks. (a) XQuAD shows consistent separation between configurations, with MONOWEB underperforming during training. (b) MLQA exhibits overlapping trajectories across all configurations. (c) HellaSwag performance across language pairs under identical FINEWEB setup shows variation, indicating cross-lingual transfer varies by pair. (d) HellaSwag within en-fr: stable performance across bilingual configurations.

alignment provided by parallel data. The latter relies more on within-language comprehension, which monolingual pretraining appears sufficient to support. Together with the translation results, this suggests a task-dependent sensitivity to bilingual data that correlates with the degree of cross-lingual generation required: tasks demanding output in a non-input language are most affected, while those relying primarily on comprehension remain robust.

Understanding and Reasoning Tasks Table 8 presents results across five understanding and reasoning benchmarks. All tasks show stability across all bilingual configurations, with performance consistently being within 1-2% of the baseline.

To better understand this stability, we take HellaSwag as a representative case. Figures 2c, and 2d demonstrate two complementary findings: First, Figure 2c compares different language pairs under the FINEWEB setting where all three pairs use identical English data for balanced training, and shows discernible variation in HellaSwag_En performance, indicating cross-lingual transfer effects exist and vary across language pairs. Second, Figure 2d compares different bilingual settings for the EN–FR pair. Performance remains unchanged whether bilingual data is present (FINEWEB), absent (MONOWEB), or partially restored (MWB+P, MWB+CS), demonstrating that cross-lingual transfer persists without bilingual data. Similar patterns also emerge across other benchmarks.

Task	FWB	MWB	MWB+P	MWB+CS
English (Avg)				
XNLI	46.3	45.6	45.9	46.8
HellaSwag	39.1	39.4	39.6	39.6
ARC_C	32.3	33.6	34.7	33.8
ARC_E	68.5	68.3	68.3	67.9
PAWS	54.5	54.9	54.0	55.5
TruthfulQA	22.0	21.8	22.4	20.9
Xwinograd	75.7	75.4	74.0	73.6
Xstorycloze	64.6	65.2	64.1	65.6
German				
XNLI	44.5	43.4	43.8	41.4
HellaSwag	34.8	35.0	35.5	35.2
ARC	22.9	24.1	24.9	25.2
PAWS	51.9	52.0	51.6	51.8
TruthfulQA	23.4	21.4	21.3	24.1
Spanish				
XNLI	43.5	42.3	43.9	45.4
HellaSwag	38.6	38.6	38.5	39.2
ARC	28.6	29.7	27.9	28.5
PAWS	50.1	53.1	51.3	51.8
TruthfulQA	25.6	26.7	25.2	26.7
Xstorycloze	62.3	61.6	61.6	61.4
French				
XNLI	44.6	44.0	44.0	44.5
HellaSwag	38.0	38.5	38.4	37.9
ARC	29.1	26.4	26.5	26.9
PAWS	52.6	47.9	52.2	53.8
TruthfulQA	24.4	25.8	22.9	25.4
Xwinograd	61.5	66.3	61.5	60.2

Table 8: Multilingual understanding and reasoning performance across all language pairs. For English, the reported numbers are averaged over three language pairs.

5.4 Explaining the Asymmetry: Why MT Collapses but Reasoning Persists

Removing bilingual data causes a severe collapse in machine translation, while cross-lingual reasoning and understanding tasks remain largely unaffected. To explain this phenomenon, we analyze how bilingual data removal impacts cross-lingual alignment at different linguistic granularities. Specifically, we measure alignment across layers using Precision@1 (P@1) computed with cosine similarity for sentence representations (3,000 WMT parallel sentences) and word representations (2,000 MUSE (Conneau et al., 2017) pairs). As shown in Table 9, we observe a stark divergence: while MonoWeb preserves robust sentence-level alignment ($< 2\%$ drop from FineWeb), it suffers a sharp 13–21% degradation in lexical-level alignment. This suggests that monolingual pretraining is sufficient to align sentence-level semantics, sup-

	Sentence-level P@1			Lexical-level P@1		
Layer	**FWB**	**MWB**	**Δ**	**FWB**	**MWB**	**Δ**
0	1.7	1.8	+0.1	5.8	8.3	+2.5
6	60.6	55.9	-4.7	40.7	19.7	**-21.0**
12	93.7	92.5	-1.3	68.7	55.1	-13.6
23	81.2	79.4	-1.8	25.5	18.4	-7.1

Table 9: Layer-wise Alignment Analysis. Lexical-level alignment shows a sharp drop at middle layers in MONOWEB, while sentence-level alignment remains largely stable.

Direction	chrF ↑		TER ↓	
	FWB	MWB	FWB	MWB
en→ar	10.95	5.91	121.5	132.6
ar→en	0.36	0.33	115.4	121.8

Table 10: Translation results on FLORES-200 for English–Arabic. We report chrF and TER as they are more robust to Arabic's rich morphology than BLEU. Results are preliminary (∼20B training tokens).

porting cross-lingual understanding and reasoning, but fails to establish the fine-grained lexical correspondences required for accurate translation.

5.5 Extension to Non-Latin Script Languages

Our main experiments focus on languages within the Latin script family. To examine whether the observed patterns generalise to typologically distant languages, we extend our analysis to Arabic (`arb_Arab`), a right-to-left, morphologically rich language with a non-Latin script. We construct the Arabic corpus by sampling 32B tokens from FINEWEB2 and apply the same two-stage bilingual document identification pipeline described in subsection 3.1. The filtering process removes 3.54% of Arabic documents, resulting in a MONOWEB variant of 30B tokens. Due to computational constraints, we evaluate only the FINEWEB and MONOWEB configurations at approximately 20B training tokens. We use chrF (Popović, 2015) as the primary evaluation metric, given its robustness to the rich morphology of Arabic.

As shown in Table 10, removing bilingual documents causes en→ar chrF to drop from 10.95 to 5.91 (46% relative decline), consistent in magnitude with the degradation observed for Latin-script pairs. The ar→en direction remains near zero for both configurations, likely because 20B training tokens is insufficient for the model to acquire Arabic generation capability at this stage. Despite being preliminary, these results suggest that the importance of bilingual data for translation generalises beyond the Latin script family.

6 Conclusion

This study explored the role of bilingual data in multilingual LLM pretraining and uncovered a clear task asymmetry. Translation is highly sensitive to a small fraction of bilingual content (2%), whereas other cross-lingual tasks remain largely unaffected. Further analysis shows that parallel data, not code-switching text, drives translation performance. This indicates that explicit cross-lingual alignment is essential for translation, while monolingual exposure largely suffices for broader cross-lingual understanding. These findings imply that multilingual pretraining may benefit more from high-quality parallel data than from large quantities of code-switching text. More broadly, our results highlight that the impact of bilingual data during multilingual pretraining can vary substantially across tasks, suggesting that its role is nuanced even within the pretraining stage.

Limitations

Our study has three limitations. First, we only pretrained 1.35B-parameter models; larger models may exhibit different sensitivity to bilingual data. Second, our experiments focus on Latin-script languages; while preliminary Arabic results suggest generalisability, whether the full pattern of task asymmetry holds for typologically distant languages remains open. Third, our bilingual data taxonomy is coarse-grained; finer distinctions such as domain, register, or alignment quality may further influence cross-lingual learning.

Acknowledgements

We acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR), which is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via the UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. We also gratefully acknowledge NVIDIA Corporation for their support through the NVIDIA Academic Grant Program, which provided GPU computing resources used in this research.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2024. Can llms really learn to translate a low-resource language from one grammar book? *arXiv preprint arXiv:2409.19151*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ales Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Eleftheria Briakou, Colin Cherry, and George F. Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability. In *Annual Meeting of the Association for Computational Linguistics*.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguistics*, 19:263–311.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries. *ArXiv*, abs/2010.12566.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and M. Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *North American Chapter of the Association for Computational Linguistics*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *ArXiv*, abs/1803.05457.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Conference on Empirical Methods in Natural Language Processing*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The Llama 3 herd of models. *ArXiv*, abs/2407.21783.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey

Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjan Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Z. Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, and 2 others. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. CoSDA-ML: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. *ArXiv*, abs/2006.06402.

Muhammad Reza Qorib, Junyi Li, and Hwee Tou Ng. 2025. Just go parallel: Improving the multilingual capabilities of large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

NLLB team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *Preprint*, arXiv:2106.12066.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cris tian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Investigating and scaling up code-switching for multilingual language model pre-training. In *Annual Meeting of the Association for Computational Linguistics*.

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A Smith, and Hannaneh Hajishirzi. 2025. Infini-gram mini: Exact n-gram search at the internet scale with fm-index. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24955–24980.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei

Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2024. Code-switching curriculum learning for multilingual transfer in LLMs. In *Annual Meeting of the Association for Computational Linguistics*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*.